Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft

Dissertation

Entwicklung einer Analysemethode für Institutional Repositories unter Verwendung von Nutzungsdaten

zur Erlangung des akademischen Grades
Doctor philosophiae (Dr. phil.)

eingereicht an der
Philosophischen Fakultät I

von
Sabine Henneberger

Präsident der Humboldt-Universität: Prof. Dr. Jan-Hendrik Olbertz

Dekan: Prof. Michael Seadle, PhD

Gutachter: 1. Prof. Dr. Peter Schirmbacher

2. Prof. Dr. Stefan Gradmann

Datum der Einreichung: 25.5.2011

Datum der Disputation: 13.10.2011

Inhalt

Abstract (Deutsch)

Nutzungsdaten von elektronischen wissenschaftlichen Publikationen und insbesondere die Anzahl ihrer Downloads rücken mit der Verbreitung des Internets zunehmend in den Blickpunkt des Interesses der Autoren, der Herausgeber, der technischen Anbieter und der Nutzer solcher Publikationen. Downloadzahlen von Publikationen, welche durch Auswertung der Protokolle der IT-Systeme der Anbieter ermittelt werden, sind solche Nutzungsdaten. Die Erhebung erfolgt durch Filterung aller stattgefundenen Zugriffe und Summierung über eine definierte Zeiteinheit.

Downloadzahlen sind Gegenstand wissenschaftlicher Untersuchungen, in welchen das Konzept des Citation Impact auf die Nutzungshäufigkeit einer Publikation übertragen und der sogenannte Download Impact gebildet wird. Besonderes Augenmerk wird dem Zusammenhang von Citation Impact und Download Impact gewidmet. Handelt es sich um Open-Access-Publikationen, muss davon ausgegangen werden, dass in den Downloadzahlen nicht nur menschliche, sondern auch maschinelle Zugriffe erfasst wurden, da eine sichere Unterscheidung unmöglich ist. Das hat zur Folge, dass die gewonnenen Daten für die einzelnen Publikationen unzuverlässig sind und starken Schwankungen unterliegen. Trotzdem enthalten sie wertvolle Informationen, welche mit Hilfe der Mathematischen Statistik nutzbar gemacht werden können.

Mit nichtparametrischen Methoden ausgewertet, geben Downloadzahlen Auskunft über die Sichtbarkeit von elektronischen Publikationen im Internet. Diese Methoden bilden den Kern von NoRA (Non-parametric Repository Analysis), mit deren Hilfe die Betreiber von Open Access Repositories die Downloadzahlen ihrer elektronischen Publikationen auswerten können, um Sichtbarkeitsdefizite zu ermitteln und zu beheben und so die Qualität ihres Online-Angebotes zu erhöhen.

Die Analysemethode NoRA wurde auf die Daten von vier universitären Institutional Repositories erfolgreich angewendet. Es konnten jeweils Gruppen von Publikationen identifiziert werden, die sich hinsichtlich ihrer Nutzung signifikant unterscheiden. Die Parallelen in den Ergebnissen weisen auf Einflussfaktoren für die Nutzungsdaten hin, welche in der gegenwärtigen Diskussion bisher keine Berücksichtigung finden. Hier erschließen sich weitere Anwendungsfelder für NoRA. Gleichzeitig geben die Ergebnisse Anlass, den Informationsgehalt von Downloadzahlen für die einzelne Publikation kritisch zu hinterfragen.

Schlagwörter:

Open Access, Citation Impact, Download Impact, Institutional Repository, Nutzungsdaten, Sichtbarkeit, Nichtparametrische Methoden, Signifikanztest

Abstract (Englisch)

With the spread of internet usage over the past decades, access characteristics of electronic scientific publications, especially the number of document downloads, are of increasing interest to the authors, publishers, technical providers and users of such publications. These download data of publications are usually obtained from the protocols of the IT systems of the provider. A data set is then created by filtering all accesses and subsequent summarizing over a certain time unit.

Download data are the subject of scientific investigations, in which the concept of the Citation Impact is applied to the rate of use of a publication and the so-called Download Impact is formed. Special attention is paid to the relation between Citation Impact and Download Impact. In the case of Open Access publications, two types of access need to be distinguished. Human access and machine access are both captured and a reliable distinction is not possible yet. As a result, the data obtained for single publications are unreliable and subject to strong fluctuations. Nevertheless, they contain valuable information that can be made useful with the help of mathematical statistics.

Analyzed with nonparametric methods, download data give information about the visibility of electronic publications on the Internet. These methods form the core of NoRA (Non-parametric Repository Analysis). With the help of NoRA, the operators of Open Access Repositories are able to analyze the download data of their electronic publications, to identify and correct deficiencies of visibility and to increase the quality of their online platform.

The analytical method NoRA was successfully applied to data from Institutional Repositories of four universities. In each case, groups of publications were identified that differed significantly in their usage. Similarities in the results reveal factors that influence the usage data, which have not been taken into account previously. The presented results imply further applications of NoRA but also raise doubts about the value of download data of single publications.

Keywords:

Open Access, Citation Impact, Download Impact, Institutional Repository, usage data, visibility, non-parametric method, test of significance

Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, ohne deren Mitwirkung die vorliegende Dissertation nicht entstanden wäre.

Besonderer Dank gilt meinem Betreuer Prof. Dr. Peter Schirmbacher, der mir die Möglichkeit für dieses Promotionsvorhaben bot und mich dabei stets bestärkte und unterstützte. Seine wertvollen Anregungen und Hinweise, die er mir in zahlreichen Diskussionen gab, trugen maßgeblich zum Gelingen der Arbeit bei. Prof. Dr. Stefan Gradmann danke ich für die Bereitschaft, das Zweitgutachten zu übernehmen.

Die sachkundige Kritik von Dr. Uwe Müller half bei der strukturellen Verbesserung der Arbeit. Durch Befolgen seiner zahlreichen Tips gelang es mir, viele Passagen prägnanter zu formulieren. Bei der Auffrischung meiner Kenntnisse der Mathematischen Statistik erhielt ich wertvolle Unterstützung durch Dr. Wolfgang Kössler. Mit aktueller Software stattete mich Wolf Lesener aus, der mir mit Rat und Tat bei deren Installation und Anwendung zur Seite stand.

Einen wichtigen Beitrag leisteten alle diejenigen, die sich bereit fanden, Daten zur Verfügung zu stellen. Das waren Annette Maile, Ulrike Fälsch, Leonhard Maylein, Florian Heß und Ulrich Herb. Bei der Akquisition weiterer Daten engagierten sich Sünje Dallmeier-Tiessen und Heinz Pampel.

Bedanken möchte ich mich vor allem bei den Mitgliedern, auch den ehemaligen, der Arbeitsgruppe Elektronisches Publizieren des Computer- und Medienservice der Humboldt-Universität unter der Leitung von Susanne Dobratz. Die Tätigkeit in dieser Arbeitsgruppe und das offene innovative Umfeld dort gaben die Anregung zur Dissertation über ein Thema, welches eng mit deren Aufgaben verbunden ist.

Für die Motivation, vor allem aber für die Unterstützung in der letzten Phase der Arbeit, danke ich meiner Familie.

Einleitung und Ziel

Wissenschaft kann ihren Zweck nur erfüllen, wenn gewonnene Erkenntnisse verbreitet, gesichert und ausgetauscht werden. Das Mittel dazu ist die Publikation von Forschungsergebnissen in den unterschiedlichsten Formen. Durch die Publikation machen Wissenschaftler die Ergebnisse ihrer Arbeit in der Fachwelt bekannt und geben dadurch gleichzeitig ein Zeugnis ihrer Produktivität und der Qualität ihrer Arbeit ab. Einmal gewonnene Erkenntnisse werden in einer für die wissenschaftliche Öffentlichkeit nachvollziehbaren Form dargestellt und aufbewahrt. Durch die Publikation ist es möglich, wissenschaftliche Communities zu bilden, die über einen gemeinsamen Kenntnisstand verfügen und über institutionelle und Ländergrenzen hinausgehen.

Mit der weltweiten Verbreitung des Internets lösen digitale Publikationen mehr und mehr die traditionelle Papierform ab. Benötigte Informationen können mit wesentlich geringerem zeitlichen Aufwand und mit größerer Vollständigkeit, als es die Papierform erlaubte, am Computer beschafft werden. Wie zuvor in Papierform sind nun in elektronischer Form veröffentlichte Artikel von Fachzeitschriften, sieht man von Fachgebieten ab, die nach wie vor das Buch bevorzugen, eine der am häufigsten genutzten Informationsquellen. Diese sind allerdings nicht generell frei verfügbar und die Nutzung verursacht in der Regel Kosten. Die Verbreitung der wissenschaftlichen Erkenntnisse wird so durch finanzielle Barrieren begrenzt.

Das Ziel des Publizierens nach dem Prinzip des Open Access ist die ungehinderte Verbreitung wissenschaftlicher Erkenntnisse. Zu diesem Zweck bieten Forschungseinrichtungen Arbeitsergebnisse von Mitarbeitern online auf sogenannten Open Access Repositories an. Auf diese Informationen kann ohne Einschränkung zugegriffen werden. Aus mehreren Gründen, die später erläutert werden, ist jedoch die Inanspruchnahme dieser Repositories als Publikationsmedium für Wissenschaftler nicht immer attraktiv. Die Betreiber versuchen deshalb, mit verschiedenen Maßnahmen die Attraktivität ihrer Repositories zu erhöhen. Eine davon ist die technisch leicht zu erzeugende Angabe, wie oft auf eine Publikation zugegriffen wurde. Sowohl Autoren als auch Rezipienten sollen dadurch über die quantitative Nutzung der Publikation informiert werden. Solche Zugriffszahlen werden bereits von vielen Repositories erhoben und angeboten. Nutzungsdaten elektronischer Publikationen sind Gegenstand wissenschaftlicher Untersuchungen, bei denen oft im Vordergrund steht, wie sich eine Publikation nach dem Open-Access-Prinzip auf deren spätere Zitation auswirkt. Die Erkenntnis, dass Open Access im Vergleich zu Non Open Access die Nutzung von Publikationen erhöht, gilt inzwischen als gesichert.

Aufbau und Betrieb eines Open Access Repository erfordern einen nicht zu unterschätzenden finanziellen und personellen Aufwand, der sich umso mehr lohnt, je besser es seine Aufgaben erfüllt. Eine dieser Aufgaben, und aus der Sicht von Autoren und Herausgebern sicher die wichtigste, ist es, Bedingungen für eine möglichst gute Verbreitung der Publikationen im Internet und damit für deren Nutzung zu schaffen. Eine dafür notwendige Voraussetzung ist die gute Sichtbarkeit des Repository im Internet. Sichtbarkeit ist zwar keine hinreichende Voraussetzung, eine geringe Nutzung ist jedoch ein Hinweis darauf, dass Sichtbarkeitsdefizite bestehen, wogegen hohe Nutzung für eine gute Sichtbarkeit spricht. Wie können Betreiber aber erfahren, wie gut die von ihnen veröffentlichten Publikationen im Internet sichtbar sind? Die Zugriffszahl auf eine

einzelne Publikation mag für die Autoren von Interesse sein, für den Betreiber sind dagegen Aspekte von Relevanz, die die Struktur des Repository als Ganzes betreffen. Ein wichtiger Gesichtspunkt ist dabei, ob nach inhaltlichen oder formalen Merkmalen gebildete Publikationsgruppen signifikante Unterschiede in ihrer Nutzung aufweisen. Eine schwächere Nutzung von beispielsweise naturwissenschaftlichen im Vergleich zu geisteswissenschaftlichen Arbeiten würde auf Defizite bei der Darstellung der ersten Gruppe hinweisen und Maßnahmen zur Beseitigung dieser Defizite erfordern. Wie allerdings solche Unterschiede verlässlich aufgedeckt werden können, ist bisher kaum untersucht.

Dieses Thema ist Gegenstand der vorgelegten Arbeit. Im Mittelpunkt steht, wie Informationen über die Nutzung von Open Access Repositories analysiert und bewertet werden müssen, um Schlussfolgerungen für die Sichtbarkeit seiner Publikationen und Konsequenzen für dessen Gestaltung ziehen zu können. Ziel ist die Entwicklung einer Analysemethode der Meta- und Nutzungsdaten eines Repository, die diese Aufgabe erfüllt. Dabei ist wesentlich, dass Zugriffszahlen starke Schwankungen aufweisen und daher einfache Mittelwertbildungen versagen. Eine der Ursachen sind maschinelle Zugriffe, die von den Betreibern nur zum Teil als solche erkannt werden. Ebenso besitzen sehr hohe Zugriffe auf einzelne Publikationen keine Aussagekraft für die Sichtbarkeit des gesamten Repository. Zur Analyse der Nutzungsdaten müssen daher adäquate Methoden der Mathematischen Statistik eingesetzt werden. Dabei erweisen sich nichtparametrische Verfahren, insbesondere der Test von Kruskal-Wallis[1], als geeignet. Dieser Test erlaubt es, in dem komplexen Datenmaterial genau solche Publikationsgruppen zu identifizieren, die sich in ihrer Nutzung signifikant unterscheiden. Zur Abkürzung wird im Weiteren das Akronym NoRA (Non-parametric Repository Aanalysis) für diese Methode verwendet.

Neben ihrer mathematisch-statistischen Eignung soll die Analysemethode weitere Kriterien erfüllen: Sie soll von den Betreibern selbst angewendet werden können, möglichst wenig Einarbeitungsaufwand erfordern und deshalb auf verbreitete Software-Werkzeuge zurückgreifen. NoRA wird außerdem in einer Form entwickelt, die die Anwendung auf viele Open Access Repositories ermöglicht, andererseits aber genug Flexibilität bietet, um individuelle Bedürfnisse und Probleme berücksichtigen zu können. Die Anwendung von NoRA erfolgt, wie bereits oben erwähnt, mit dem Ziel, Betreibern Informationen über die Entwicklung und Nutzung des Repository zur Verfügung zu stellen und Möglichkeiten und Notwendigkeiten der Verbesserung aufzuzeigen. Verbesserungen sind vor allem dann notwendig, wenn die Analyse ergibt, dass aufgrund geringer Nutzung Sichtbarkeitsnachteile für Gruppen von Publikationen, die einen wesentlichen Bestandteil des Repository bilden, vermutet werden müssen. Damit unterstützt die Methode die Betreiber bei der Erfüllung ihrer Aufgabe, den Service für Autoren, Herausgeber und Rezipienten zu verbessern und damit die Attraktivität und Akzeptanz des Repository zu erhöhen.

[1] Der Test ist nach William Kruskal und Wilson Allen Wallis benannt.

Die Analysemethode wurde anhand der Daten von vier Repositories von Universitäten entwickelt und erprobt. Deshalb wurden für Universitäten gebräuchliche und in den Metadaten enthaltene Publikationstypen und Strukturen abgebildet. Dadurch erfolgt die Darstellung der Analysemethode in einer Form, die spezielle Anforderungen von Universitäten berücksichtigt. Das Prinzip von NoRA ist jedoch auf jede Art von Open Access Repository anwendbar und kann für die Beantwortung verschiedenster Fragestellungen im Zusammenhang mit der Nutzung dienen. Damit entstand im Ergebnis eine sehr flexibel einsetzbare Analysemethode. Daneben zeigten die Ergebnisse der vier Repositories, dass es ein für Institutional Repositories typisches Nutzerverhalten gibt, welches bei der Gestaltung der Websites beachtet werden muss, um Benachteiligungen für Publikationen hinsichtlich ihrer Sichtbarkeit gering zu halten.

Im 1. Kapitel wird der Leser mit den Hintergründen vertraut gemacht, vor welchen die Analysemethode entwickelt wurde und in welche Zusammenhänge sich das Ziel der Arbeit einordnet. In der Arbeit verwendete Begriffe werden eingeführt und erläutert, soweit das für das Verständnis notwendig ist. Auf weiterführende und ausführlichere Behandlung der einzelnen Themen wird jeweils hingewiesen. Die Gründe, warum der Impact wissenschaftlicher Publikationen so enorm wichtig für Autoren ist, werden aufgeführt. Es wird erklärt, wie der Impact quantifiziert und das Konzept des Citation Impact auf die Nutzung der Publikation in Form des Download Impact übertragen wird, da sich daraus erklärt, warum dem Download Impact und damit der Nutzungshäufigkeit elektronischer Publikationen große Bedeutung zugemessen wird. Die kritische Auseinandersetzung mit der Anwendungspraxis des Citation Impact soll auf Probleme aufmerksam machen, die durch Überbewertung und falsche Interpretation von Impact-Maßen von Publikationen entstehen und ebenso im Umgang mit dem Download Impact auftreten können. Des Weiteren erfolgt eine Darstellung der Ziele der Open Access-Bewegung und der Funktion der Institutional Repositories innerhalb dieser Bewegung. Aus diesem Zusammenhang lassen sich eine Reihe von Aufgaben der Institutional Repositories und Kriterien, die sie erfüllen müssen, ableiten. Ihre Funktion in der Institution schließt diese Aufgaben ein, geht aber weit darüber hinaus. Welchen Qualitätsstandards sie genügen sollten und welche formalen Voraussetzungen dafür erfüllt sein müssen, wird erläutert.

Das 2. Kapitel widmet sich ausführlich der Ermittlung von Nutzungsdaten in Form von Downloadzahlen aus den Logfiles von Webservern. Dabei wird deutlich gemacht, welchen Einflüssen Downloadzahlen unterliegen und inwieweit eine Unterscheidung zwischen menschlicher Nutzung und maschinellem Zugriff möglich ist. Anhand des Vergleichs von fünf Tools zur Logfileanalyse wird nachgewiesen, dass Downloadzahlen durch unterschiedliche Algorithmen, Konfigurationen der Tools und Filtermethoden beeinflusst werden. Die Konsequenz daraus ist, dass Downloadzahlen von Publikationen unter identischen Bedingungen erhoben werden müssen, damit sie vergleichbar sind. Des Weiteren wird nachgewiesen, dass es unmöglich ist, alle maschinellen Zugriffe als solche zu identifizieren, eine Tatsache, die bei der Analyse von Downloadzahlen zu berücksichtigen ist.

Gegenstand des 3. Kapitels ist die Entwicklung der Analysemethode. Hier wird genauer erläutert, mit welchem Ziel die Analysemethode angewendet werden kann und welchem Prinzip dabei gefolgt wird. Danach wird dargestellt, wie bei der Entwicklung der Analysemethode vorgegangen wurde. Die Unterschiede und

Gemeinsamkeiten des vorhandenen Datenmaterials werden beschrieben und gemeinsame Merkmale wie Publikationstyp und inhaltliche Klassifikation ermittelt, um für alle vier Repositories eine einheitliche Form der Analyse durchführen zu können. Anschließend wird eine Datenstruktur entwickelt, die alle für die Analyse verwendeten Daten enthält. Anhand von Beispielen werden charakteristische Eigenschaften der Daten demonstriert und aufgrund dieser Eigenschaften die statistischen Verfahren ausgewählt, die zur Analyse der Nutzungsdaten geeignet sind.

Im folgenden 4. Kapitel wird die Analysemethode in sechs Schritten, die nacheinander an einem Beispiel ausgeführt werden, in übersichtlicher Form beschrieben und demonstriert. In diesen sechs Schritten ist sowohl die Analyse der Metadaten als auch die Nutzungsdatenanalyse und die grafische Darstellung der Ergebnisse enthalten. Die Analysemethode wurde auf die Daten der vier Repositories in einheitlicher Form angewendet. Es ergaben sich in allen vier Fällen signifikante Unterschiede der Downloadzahlen für Gruppen von Publikationen, wobei sich die Zusammensetzung der Gruppen von Fall zu Fall unterscheidet. Die Ergebnisse und ihre Darstellungen sind Inhalt des 5. Kapitels. Für jedes der vier Repositories werden die Ergebnisse zusammengefasst und diskutiert und Hinweise gegeben, welche Gruppen von Publikationen aufgrund ihrer Downloadzahlen und der Anzahl der enthaltenen Publikationen besondere Beachtung finden müssen. Es wird gezeigt, dass sich Downloadzahlen trotz bekannter Probleme bei der Erhebung für eine Analyse eignen und die entwickelte Analysemethode verwertbare Ergebnisse liefert.

Im 6. Kapitel erfolgt zunächst eine Bewertung von NoRA, die auf der vorher dargestellten Anwendung auf vier Repositories basiert. Anschließend werden die Ergebnisse der Analysen diskutiert. Dabei wird nicht auf die absoluten Unterschiede der Downloadzahlen der vier Repositories eingegangen, da diese unter verschiedenen Bedingungen erhoben wurden und demzufolge nicht vergleichbar sind. Die Ergebnisse in Form der Relationen der Downloadzahlen zeigen jedoch auffällige Gemeinsamkeiten, die darauf hinweisen, dass es für Institutional Repositories ein typisches Nutzerverhalten gibt, welches berücksichtigt werden muss. Die relativen Unterschiede der Downloadzahlen lassen sich zum Teil erklären, wenn man die Websites der Repositories analysiert und Beziehungen zu den Downloadzahlen herstellt. Hier wird deutlich, welchen Einfluss bereits die Struktur der Website auf die Sichtbarkeit und die spätere Nutzung der Publikationen hat. Diese Erkenntnis stimmt mit den Ergebnissen anderer Untersuchungen zur Sichtbarkeit im Internet überein und führt zu Schlussfolgerungen für die Bewertung und Verwendung von Downloadzahlen.

Als Abschluss erfolgt eine Zusammenfassung der Arbeit. Es wird auf die Konsequenzen aufmerksam gemacht, die sich aus der Analyse der vier Institutional Repositories für die Interpretation von Downloads und den Umgang mit Nutzungsstatistiken ergeben. Im Ausblick wird auf die Möglichkeit hingewiesen, dass das Konzept von NoRA auf die Analyse von Repositories, die in einem Netzwerk verbunden sind, übertragen werden kann. Unter der Voraussetzung standardisiert erhobener und zentral aggregierter Nutzungsdaten ist es mit den Prinzipien von NoRA möglich, ein Ranking von Open Access Repositories durchzuführen, welches deren Nutzung berücksichtigt.

1 Hintergründe, Zusammenhänge und Begriffe

Um zu verstehen, warum Nutzungsdaten für wissenschaftliche elektronische Publikationen erhoben werden und wieso das Interesse dafür so groß ist, muss genauer auf einige Aspekte des Wissenschaftsbetriebes eingegangen werden, die nicht isoliert betrachtet werden können, sondern eng zusammenhängen und sich gegenseitig bedingen. In diesem Kapitel werden die Hintergründe und Zusammenhänge beleuchtet und die wichtigsten Begriffe eingeführt und erklärt. An dieser Stelle kann keine Darstellung geboten werden, die auch nur entfernt Anspruch auf Vollständigkeit hat, da Themen angesprochen werden, die Gegenstand verschiedener Wissenschaftsdisziplinen wie Philosophie, Sozial- und Informationswissenschaften, Informatik und Mathematik sind. Die Themen werden soweit behandelt, wie es zum Verständnis der Dissertation nötig ist. Es wird auf umfassendere Darstellungen und Informationen hingewiesen. Speziell auf Methoden, Verfahren und Algorithmen, die im folgenden Kapitel genannt werden, kann nur in sehr allgemeiner Form eingegangen werden.

Ein weiteres Ziel dieses Kapitels ist es, den Leser für die Probleme der Bewertung von wissenschaftlichen Publikationen in Form von Maßen zu sensibilisieren. Der Zitationsanalyse und der Kritik daran wurde deshalb ein relativ breiter Raum eingeräumt. Bei der Übertragung des Konzeptes des Citation Impact auf die Nutzung potenzieren sich diese Probleme, so dass der Umgang mit Nutzungsdaten für eine Publikation größte Vorsicht erfordert.

1.1 Publizieren in der Wissenschaft

Die Publikation[2] von Ergebnissen wissenschaftlicher Arbeit in Form von Büchern, Zeitschriften usw. ist ein Vorgang, ohne den es nicht möglich wäre, die Ziele der Wissenschaft, wie sie von Robert Merton festgestellt wurden, zu erreichen: „ Das institutionelle Ziel der Wissenschaft ist die Ausweitung gesicherten Wissens. Die technischen Methoden, die zur Erreichung dieses Zieles angewandt werden, liefern die relevante Definition von Wissen: empirisch bestätigte und logisch konsistente Voraussagen.“ (Merton 1972, S. 47).

Nicht nur die Ausweitung des Wissens bedarf der Kommunikation der Wissenschaftler untereinander, die Wissenschaft selbst wird als Kommunikationssystem betrachtet: „Das Wissenschaftssystem ist ein Kommunikationssystem, in dem die Ergebnisse der Forschung zwischen den Mitgliedern der jeweiligen scientific communities kommuniziert und der kollegialen Kritik unterworfen werden.“ (Weingart 2003, S. 32). Nach der Theorie der wissenschaftlichen Kommunikation von Michailow, Cernyi und Gilarewskij (Michailow et al. 1970) kann zwischen informeller und formeller Kommunikation unterschieden werden. Die informelle Kommunikation findet vor allem über persönliche Kontakte und am effektivsten durch Gespräche statt, wo-

[2] Ausführliche Definitionen, Begriffsbestimmungen und Erläuterungen zu den Themen Publikation und Wissenschaftliches Publizieren findet man in (Müller 2009; Stock 1998).

bei die ausgetauschten Informationen einem begrenzten Kreis von Wissenschaftlern zugänglich sind. Die formelle Kommunikation beruht dagegen auf der Nutzung von allgemein zugänglichen Publikationen (Stock 1998).

Die Verbreitung von Resultaten geschieht zum überwiegenden Teil über den Weg der formellen Kommunikation durch Publikationen. Über Verlage werden Publikationen, die in der Regel einem Peer Review unterzogen wurden, einer wissenschaftlichen Community[3] zur Kenntnis gebracht.

Merton weiter: „Der Zwang zur Verbreitung von Resultaten wird durch das institutionelle Ziel der Erweiterung des Wissens sowie durch den Anreiz der Anerkennung verstärkt, die natürlich von der Veröffentlichung abhängt." (Merton 1972, S. 51). Durch Publikation erst wird die intellektuelle Urheberschaft an einem Forschungsergebnis auch außerhalb des engsten fachlichen Umfeldes bekannt gemacht und die Leistung kann öffentlich wahrgenommen und anerkannt werden. Auf diese Wahrnehmung und Anerkennung, nicht nur durch Peers, sondern auch durch Institutionen, ist ein einzelner Wissenschaftler oder auch eine wissenschaftliche Institution angewiesen, um Reputation zu erlangen und dadurch die Akquisition von Ressourcen zu ermöglichen. Stefan Hornbostel schreibt dazu, „dass das Publikationswesen für die Zuteilung von Anerkennung, für die Bewertung von Leistungen, für die Anerkennung von Prioritäten und Geltungsansprüchen und für die inhaltliche Bestimmung des Forschungsfeldes entscheidende Bedeutung hat" (Hornbostel 1997, S. 238).

Beschaffung von Publikationen anderer Autoren und Publikation von eigenen Resultaten sind integrale Bestandteile wissenschaftlicher Arbeit. Noch bis vor kurzem war es für Wissenschaftler der übliche Weg, Publikationen auf Papier über Bibliotheken, von Verlagen oder direkt vom Autor zu beziehen. Diese kursierten in Instituten und Arbeitsgruppen und erst modernere Methoden erlaubten es, Papierkopien auf bequemen Weg herzustellen. Die Beschaffung war mühsam und kostete Zeit. Mit Beginn der 1990er Jahre wurden Publikationen zunehmend auf elektronischem Weg erstellt und verbreitet. Die Verlage gingen dazu über, sowohl Papier- als auch elektronische Versionen anzubieten und die Tendenz geht in Richtung der Ablösung der Papierform durch das elektronische Format. Das Angebot an Publikationen vergrößerte sich insgesamt und auch individuell für den einzelnen Wissenschaftler. Inzwischen ist das Beziehen einer digitalen Kopie zur vorherrschenden Methode bei der Informationsbeschaffung geworden. Das elektronische Publizieren führte durch vorher nicht gekannte Recherche- und Vernetzungsmöglichkeiten bis hin zur Verfügbarkeit von Primärdaten zu völlig neuen Qualitäten der formalen Kommunikation. Gleichzeitig wurden durch das Internet viele neue Möglichkeiten der informellen Kommunikation (z. B. Online-Foren, Blogs) geschaffen und die Beteiligung daran wird kaum durch technische Hürden beeinträchtigt.

Die Beschleunigung des Wissenschaftsprozesses führt zur kontinuierlichen Vergrößerung der Anzahl von Publikationen[4], was aber nicht bedeutete, dass die Teilhabe am wissenschaftlichen Gedankengut, obwohl in

[3] Näheres zu wissenschaftlichen Communities als Kommunikationsgemeinschaften findet man in (Weingart 2003).

technischer Hinsicht einfach möglich, im gleichen Maß gewährleistet wird. Im Gegensatz zur informellen zeichnet sich die formelle Kommunikation vor allem bei Fachzeitschriften dadurch aus, dass die Beteiligungsmöglichkeiten, obwohl technisch kein Problem, sich nicht verbessert sondern verschlechtert haben. Zeitschriftenverlage bauten bereits vorhandene Monopolstellungen aus und nutzten die Abhängigkeit des Wissenschaftsbetriebes für Preissteigerungen. Seit etwa 20 Jahren verschlechtert sich dadurch die Situation der Versorgung mit wissenschaftlicher Literatur durch Bibliotheken, egal ob traditionell auf Papier oder elektronisch. Bibliotheken sind aufgrund steigender Preise zunehmend weniger in der Lage, notwendige Publikationen zu kaufen und ihren Lesern anzubieten. Man spricht von der Serial Crisis (Swan 2006b).

Vor diesem Hintergrund wurde nach einer Alternative gesucht, die die ungehinderte Verbreitung von Publikationen und ungehinderten Zugriff darauf bietet. Wissenschaftler schlossen sich mit dem Ziel zusammen, die Diskrepanz zwischen technisch Machbarem und der vorherrschenden Praxis zu beseitigen und jedem, der über die technischen Voraussetzungen verfügt, die Teilnahme an der formellen Kommunikation der Wissenschaft zu ermöglichen. In der Folge davon formierte sich die Open Access-Bewegung.

Bevor auf Entstehung, Entwicklung und Bedeutung der Open Access-Bewegung näher eingegangen wird, sollen zuvor die im Kontext dieser Arbeit wichtigen Begriffe des Citation Impact und Download Impact eingeführt und erläutert werden.

1.2 Der Impact wissenschaftlicher Publikationen

Im Zusammenhang mit wissenschaftlichen Publikationen spricht man von Impact, wenn man die (vor allem kurzfristige) Wirkung in Fachkreisen meint. Diese, so lautet eine Annahme, wird widergespiegelt, in dem sich die Autoren auf vorangegangene Publikationen beziehen, d.h. diese zitieren. Um ein Maß für den Impact zu bilden, wird die Anzahl der Zitationen in nachfolgenden Publikationen hinzugezogen[5].

Dem kommt entgegen, dass sich Zitationen einfach durch Zählung messen lassen. Die Bedeutung, die Impact-Maße aus Zitationen als Wissenschaftsindikatoren bei der Beurteilung von Publikationen und damit von der Leistungsfähigkeit von Autoren bis heute erlangt haben, hängt mit dieser einfachen Möglichkeit, die durch elektronische Datenverarbeitung auf Datenbanken gestützt automatisch durchgeführt wird, zusammen,

[4] Wie groß die Anzahl von wissenschaftlichen Publikationen weltweit ist, kann nicht genau festgestellt werden, da die Ergebnisse von Berechnungen aufgrund verschiedener Kriterien dafür, welche Publikationen eingehen, sich stark unterscheiden. Auf der Datengrundlage von Ulrich's Web (http://www.ulrichsweb.com, gelesen 25.3.2011), einem Verzeichnis von Zeitschriften aus aller Welt, wurde eine durchschnittliche jährliche Wachstumsrate von 3,46 % für wissenschaftliche Zeitschriften seit 1800 bis zur Gegenwart berechnet (Mabe 2003).

[5] Derek de Solla Price spricht nicht nur von Wirkung, sondern sogar von Nutzen: „Wir betrachten den Nutzen einer Arbeit anhand der Häufigkeit, mit der sie zitiert wird." (de Solla Price 1974, S. 89).

andererseits aber auch mit der Annahme, dass mit Zitationen ein Belohnungssystem in der Wissenschaft realisiert wird (Hornbostel 1997, S. 284).

Zitationen können aber eine weitere Bedeutung haben. So kann man laut Cozzens grob zwei Gruppen unterscheiden: Zitation als „reward“ und Zitation als „rhetorical“[6]. Dabei sind mit „rhetorical“ Zitationen gemeint, die sich auf Publikationen beziehen, die Ergebnisse anderer Publikationen erklären, die wiederum nicht Ergebnisse des zitierten Autors sein müssen. Der so zitierte Autor wird also durch Zählung überbewertet, da sich die Zitation nicht auf seine Leistung bezieht. Zitationen können auch beiden Gruppen gleichzeitig angehören. Cozzens fordert daher einen Paradigmenwechsel in der Zitationsanalyse, nämlich die Abkehr von der Annahme, die Erzeugung von Zitationen folge immer einem „Belohnungssystem“ (Cozzens 1989). Damit wird die Annahme, die der Zitationsanalyse und ihrer Anwendung zugrunde liegt, in Frage gestellt.

Für Zitationen kann es eine Vielzahl von Anlässen geben, die sogar das Gegenteil einer Belohnung sein können und dazu dienen, sich von den Aussagen einer Publikation zu distanzieren. Die Auswahl, welche Publikationen zitiert werden, richtet sich auch häufig danach, ob diese in renommierten Zeitschriften veröffentlicht wurden und von wem sie verfasst wurden[7]. Trotz der bestehenden Zweifel, dass Zitation eine Art von Belohnung ausdrückt, die der Qualität der Publikation gilt, wird bei der Bewertung wissenschaftlicher Arbeit von dieser Annahme ausgegangen und dieser Bewertung großes Gewicht beigemessen.

Ein weiteres Konzept, die Wirkung von wissenschaftlichen Publikationen zu messen, geht davon aus, dass sich die Wirkung im Interesse an der Publikation widerspiegelt. Das Interesse wird durch Nutzung bekundet, deren Häufigkeit erfasst werden kann. Nach Derek de Solla Price ist die „Häufigkeit der Benutzung ein vernünftiges Maß für die wissenschaftliche Bedeutung einer Zeitschrift oder einer Forschungsarbeit“ (de Solla Price 1974, S. 89).

Die Messung der Nutzung von elektronischen Publikationen erfordert wenig Aufwand und wurde zuerst für digitale Bibliotheken und Artikel elektronischer kostenpflichtiger Zeitschriften durchgeführt, bei denen der Zugriff nur für angemeldete Nutzer möglich ist. Sie diente bisher vor allem der Planung in Bibliotheken und der Preisgestaltung von Verlagen, aber nicht der Bewertung von Publikationen. Diskutiert wird jedoch der Zusammenhang zwischen Häufigkeit der Nutzung und dem durch Zitationsanalyse gemessenen Impact.

[6] In der Arbeit wird noch eine dritte Gruppe genannt, Zitation als „communication“, welche aber als untergeordnet betrachtet wird.

[7] Welche weiteren Gründe zur Zitation einer Publikation führen, die nicht als Belohnung für die Qualität ihres Inhalts gelten können und deshalb den Wert von Zitationsanalysen relativieren, findet man z. B. in (Hornbostel 1997; Merton 1968; Müller 2009; Ohly 2010; Stock 1998).

1.2.1 Der Citation Impact

Maße für den Impact durch Zitationsanalysen, d.h. durch Analyse der Beziehungen von zitierenden und zitierten Publikationen, zu bilden, ist ein Teilgebiet der Bibliometrie, welche Publikationen mit quantitativen Methoden untersucht. Bibliometrie wiederum wird vielfach der Scientometrie, der Wissenschaft von der Messung der Wissenschaft, zugeordnet. Häufig werden die Begriffe Bibliometrie und Scientometrie aber synonym verwendet[8]. Die durch Zitationsanalysen errechneten Maße geben einen Citation Impact an. Dabei werden Maße zur Bewertung von ganzen Zeitschriften, einzelner Artikel und Autoren entwickelt. Eine Methode, die Qualität der entwickelten Maße zu überprüfen, ist die Untersuchung der Korrelation zwischen errechneten Maßen und der Einschätzung durch Peers[9].

Der Vorschlag, einen umfassenden Index wissenschaftlicher Publikationen und ihrer Zitationen zu erfassen, kam von Eugene Garfield (Garfield 1955), welcher seine Idee 1963 mit dem ersten Science Citation Index verwirklichte, der eine Vielzahl von Zeitschriften, vor allem aus dem naturwissenschaftlichen Bereich, abdeckte. Die Geschichte der Zitationsanalyse begann zwar wesentlich früher mit Shepards Citation Index von 1870[10], aber erst der Science Citation Index von Garfield, der zuerst als Buch und später als CD erschien, machte die Ergebnisse der Zitationsanalysen einer breiten Öffentlichkeit bekannt. Die im Index enthaltenen Daten wurden durch Zählungen von Publikationen und Zitationen ausgewertet, die zueinander ins Verhältnis gesetzt werden. Hierbei kommt es nicht darauf an, in welcher Publikation sich eine Zitation befindet. Publikationen und Zitationen als Netzwerk aufzufassen, in welchem die Publikationen die Knoten und die Zitationen die Kanten darstellen, geht auf die Idee von Derek de Solla Price zurück. Unter Annahme des Netzwerkmodells eröffnet sich die Möglichkeit, mit graphentheoretischen Methoden, wie sie auch in den Sozialwissenschaften Verwendung finden, Zitationsanalysen durchzuführen (de Solla Price 1965). Dabei ist es nicht nur von Bedeutung, wieviel Zitationen vorhanden sind, sondern auch, in welchen Publikationen sich die Zitationen befinden.

Durch Bildung spezieller Maße und Gewichtungen wird versucht, die Realität besser als mit einfachen Zählungen und Verhältniszahlen oder mit Netzwerken aus Publikationen und Zitationen abzubilden. Auch Nutzungszahlen, Umfragedaten und Kontextanalysen finden als zusätzliche Informationen Eingang. In der Praxis haben sich, sicher auch der einfachen Umsetzung wegen, die Bewertungen durchgesetzt, die nur auf der Zählung von Zitationen beruhen.

---

[8] Der Begriff „Bibliometrie“ wurde 1969 von Alan Pritchard (Pritchard 1969, zitiert in Havemann 2009) geprägt Die Bezeichnung „Scientometrie“ wurde erstmalig von Gennadij M.Dobrov (Dobrov 1974) verwendet, der sich dabei auf Arbeiten von Derek de Solla Price bezieht, der zusammen mit Dobrov zu den Begründern der Scientometrie zählt. Einen Überblick und detaillierte Informationen zum Thema „Bibliometrie“ findet man in (Havemann 2009).

[9] Ein Beispiele für diese Vorgehensweise findet man in (Bornmann et al. 2008).

1.2.2 Beispiele für Zitationsanalysen

Das bekannteste Beispiel ist die Messung des Science Citation Factor für Journale[11], auch Journal Impact Factor (JIF) genannt.

Eugene Garfield gründete 1960 das Institute for Scientific Indexing (ISI)[12]. Auf der Basis des Science Citation Index, der heute Publikationen von fast 7000 wissenschaftlichen Zeitschriften aus 150 Disziplinen ab dem Jahr 1899 enthält[13], wird der JIF berechnet. Der JIF wird seitdem nicht nur zur Bewertung von Zeitschriften, sondern auch zur Bewertung der wissenschaftlichen Leistung von Personen und Institutionen genutzt. Die Leistung wird daran gemessen, wie hoch der JIF der Zeitschrift ist, in welchem Artikel veröffentlicht wurden. Man kommt also über den Umweg der Bewertung von Zeitschriften zur Bewertung der Leistung von Personen. Dieses Vorgehen wird damit begründet, dass der Veröffentlichung eines Artikels in einer Zeitschrift eine Qualitätsprüfung durch ein Peer-Review-Verfahren vorangeht und dadurch das Niveau sichergestellt wird. Die Autoren eines Artikels, der selbst wenig Zitate erhält, profitieren von hohen Zitationszahlen anderer. Hat ein Artikel mehrere Autoren, profitieren alle Autoren unabhängig von ihrem individuellen Beitrag, wenn der Artikel in einer renommierten Zeitschrift erscheint.

Ein anderes Beispiel für die Zitationsanalyse ist der Hirsch-Index, kurz h-Index[14], der die Bewertung der Publikationen und damit der wissenschaftlichen Leistung eines Autors direkt vornimmt, ohne den Umweg über den JIF zu gehen (Hirsch 2005). Hirsch selbst empfahl, als Datengrundlage für die Berechnung die Publikationen zu verwenden, die in die Bildung des Science Citation Index einbezogen werden. Dem h-Index wird zugeschrieben, nicht nur Qualität, sondern auch Quantität der wissenschaftlichen Leistung und somit auch Produktivität[15] eines Wissenschaftlers widerzugeben (Bornmann et al. 2008). In der Folge entstanden eine Reihe von Variationen des h-Indexes, z. B. g-Index (Egghe 2006) und R-Index (Jin et al. 2007), mit denen versucht wird, bekanntgewordene Defizite des h-Index, z. B. der Bias, der durch das Alter entsteht, auszugleichen. Der h-Index wird sowohl zur Bewertung einzelner Autoren als auch ganzer Institutionen benutzt, in dem die Mittelwerte der h-Indices der Mitarbeiter gebildet und verglichen werden. Das Konzept

---

[10] In der Literatur befindet sich ein Hinweis darauf, dass es schon im Jahr 1300 einen Citation Index gegeben hat (Cronin 2001).

[11] Zahl der Zitate im laufenden Jahr auf die Artikel der vergangenen 2 Jahre / Zahl der Artikel der vergangenen 2 Jahre

[12] Das Institute for Scientific Information (ISI), wurde 1992 von der Thomson Corporation erworden und als Thomson Scientific weitergeführt. Die Thomson Corporation heißt nach der Übernahme der Nachrichtenagentur Reuters seit 2008 Thomson Reuters.Das Web of Science,gleichbedeutend zu ISI Web of Knowledge; ist eine kostenpflichtige Zitationsdatenbank, die vom ISI entwickelt wurde.

[13] Quelle: Thomson Reuters, http://thomsonreuters.com/products_services/science/science_products/a-z/science_citation_index_expanded/, gelesen 10.1.2011. Die Zahlen, die in der Literatur zu finden sind, differieren stark.

[14] Hirsch-Index h : Anzahl von Veröffentlichungen des Autors, die mindestens jeweils h Zitationen haben.

[15] Die Anzahl von Veröffentlichungen allein wird auch als Maß für die Produktivität von Autoren verwendet.

wird auch in Form des Journal Citation h-Index[16] zum Ranking von Zeitschriften eingesetzt (Braun et al. 2006). Durch gezielte Strategien der Selbstzitation ist es möglich, den h-Index zu erhöhen. Um manipulatives Verhalten von Autoren aufzudecken, wurde der q-Index definiert (Bartneck und Kokkelmans 2010).

Die bisher aufgeführten Beispiele beruhen auf Zählungen von Zitationen. Eine Anwendung des Netzwerkmodells ist dagegen der CiteRank, der mit Hilfe des PageRank-Algorithmus gebildet wird (Walker et al. 2007). Eine Kombination des JIF mit dem Ergebnis des PageRank-Algorithmus, den Y-Faktor, schlagen Bollen et al. für die Bewertung von Zeitschriften vor (Bollen et al. 2006). Nach Meinung der Autoren kann so die Bedeutung von Zeitschriften besser widergeben werden als mit einem der beiden Verfahren allein.

Inzwischen ist der ehemalige Science Citation Index eine von sieben Datenbanken des Web of Science (WoS), welches wiederum Bestandteil des Thomson Reuters Web of Knowledge ist[17]. Im Web of Knowledge wird weitaus mehr als das Ranking von Zeitschriften angeboten. Die Datenbanken enthalten ebenso Konferenzbände, Bücher, Workshops und mehr. Ausgefeilte Recherchen und Analysen von Publikationen und Autoren sind möglich. Selbstverständlich wird auch der h-Index von Autoren berechnet. Das Ranking steht nicht mehr ausschließlich im Vordergrund. Durch Auswertung von Zitaten kann durch Navigation im Zitationsnetzwerk schnell ein Überblick über die Beziehungen von Themen, Forschergruppen und Institutionen gewonnen werden. Wissenschaftler haben hiermit ein sehr komfortables Instrument zur Verfügung, welches zu einer vorher unbekannten Qualität der formellen Kommunikation geführt hat.

Blieb die Nutzung des Science Citation Index lange Zeit die einzige Möglichkeit, Rankings durchzuführen, haben sich in letzter Zeit mehrere Internetangebote etabliert, die über unterschiedliche Datenbasen verfügen.

Ebenfalls zunächst kostenpflichtig wie das Web of Knowledge, jetzt aber frei verfügbar, bietet der Verlag Elsevier seit 2004 die Abstract- und Zitationsdatenbank Scopus[18] an. Die Datenbasis wurde um Websites erweitert. Scopus gibt an, 18 000 Journale ab 1999[19] in seiner Datenbank auszuwerten und wissenschaftliche Literatur am besten abdecken zu können. In Scopus werden mit SNIP (Moed 2010) und SRJ (Gonzalez-Pereira et al. 2010)[20] Indikatoren errechnet, die kontextuelle Einflüsse einbeziehen und Rankings über Fachgebietsgrenzen hinaus ermöglichen.

Zitationen werden auch von wissenschaftlichen Suchdiensten ausgewertet. Vom Beginn 2004 an frei zugänlich ist der Suchdienst Google Scholar[21]. Welche Zeitschriften und andere Publikationsquellen genau erfasst werden, ist unbekannt. Ausgewertet werden die meisten Onlinezeitschriften, bei denen eine Qualitätskontrol-

[16] Journal Citation h-Index h: Anzahl von Artikeln in einem Journal, die mindestens jeweils h Zitationen haben.

[17] http://isiwebofknowledge.com/, gelesen 6.1.2011.

[18] Informationen: http://www.b-i-t-online.de/neues/607, gelesen 10.1.2011.

[19] http://www.info.sciverse.com/journalmetrics/search2.php, gelesen 10.1.2011, siehe auch Scopus Support.

[20] SNIP: Source Normalized Impact per Paper; SRJ: SCImago Journal Rank.

[21] http://scholar.google.de/, gelesen 11.1.2011.

le durch Peer-Review-Verfahren[22] stattfindet, die großen US-amerikanischen Wissenschaftsverlage und Onlineangebote wie Open Access Repositories. Als Ergebnis einer Suche werden Artikel in der Reihenfolge der Anzahl der Zitationen mit Links zu den zitierenden Artikeln angeboten. Google Scholar ist ein Beispiel für viele solcher Dienste, die vielfach auf bestimmte Fachgebiete spezialisiert sind und unterschiedliche Services anbieten[23].

1.2.3 Kritik an der Anwendungspraxis

Verschiedene Zitationsanalysen und Impact-Maße bilden das angenommene „Belohnungssystem" in der Wissenschaft unterschiedlich gut ab und sind daher mehr oder weniger geeignet, als Indikatoren für Leistung zu gelten. Für sich allein sind sie noch keine Qualitätsmessung, obwohl sie oft als solche wahrgenommen werden. Dobrov schreibt dazu, dass sich „…diese Häufigkeit des Zitierens als Kennziffer für die direkte Einschätzung der individuellen Ergebnisse einzelner Forscher gewöhnlich als anfechtbar und häufig als unannehmbar erwiesen (hat.)" (Dobrov 1974, S. 59). Hornbostel bezeichnet sie als eine Art „Erfolgsindikator". Und Hornbostel weiter: „Erst zusammen mit weiteren Informationen und Expertenwissen lassen sich aus den Angaben der Zitationsanalyse qualitative Einschätzungen wissenschaftlicher Leistungen entwickeln." (Hornbostel 1997, S. 307-308).

Obwohl anerkannt ist, dass die Bewertung wissenschaftlicher Publikationen oder wissenschaftlicher Leistungen durch Zitationsanalysen fragwürdig ist und sie lediglich als Hinweis dienen kann, haben ihre Ergebnisse großen Einfluss auf personelle und finanzielle Entscheidungen im Wissenschaftsbetrieb. Man meint, mit der Zitationsanalyse ein objektives Instrument zur Entscheidungsfindung in der Hand zu haben, welches subjektiven Einschätzungen, z. B. durch Peer Review, überlegen ist und vernachlässigt dabei wohlbekannte verzerrende Einflüsse.

Robert Merton wies durch Untersuchung psychosozialer Effekte den Matthäus-Effekt[24] in der Wissenschaft nach, der sich auf die Reputation von Personen und im Belohnungssystem durch einen positiven Rückkopplungseffekt auswirkt (Merton 1968). Für die Bewertung von Publikationen durch Zitationsanalysen bedeutet das, dass Publikationen oder Autoren, die bereits eine führende Position im Ranking einnehmen, ihre Position immer mehr ausbauen, da sie in Zukunft noch häufiger zitiert werden.

Eine sehr kritische Auseinandersetzung mit der Verfahrensweise, wissenschaftliche Leistungen durch Ergebnisse von Zitationsanalysen zu berurteilen, findet im Report der International Mathematical Union (IMU) statt. Anhand vieler Beispiele wird gezeigt, wie grotesk oft Rankings sind, die sich auf vermeintlich objektive

[22] Siehe dazu (Müller 2009).

[23] Eine Übersicht findet man bei Wikipedia: http://en.wikipedia.org/wiki/Academic_databases_and_search_engines, gelesen 11.1.2011.

[24] Der Matthäus-Effekt wurde nach einem Bibel-Zitat benannt:„Denn wer da hat, dem wird gegeben werden, dass er Fülle habe; wer aber nicht hat, von dem wird auch genommen, was er hat."

Ergebnisse stützen. Dabei steht außer Frage, dass Zitationsanalyse ein wichtiges Hilfsmittel sein kann, wenn sie denn sinnvoll angewendet wird. Leider, so die Autoren des Reports, ist das aber oft nicht der Fall. So kommt es in der Praxis häufig zu der Situation, dass zum einen dem Ergebnis von Zitationsanalysen ein zu großes Gewicht beigemessen wird, zum anderen diese falsch angewendet werden (Adler et al. 2008).

Trotz der breiten Nutzung des Science Citation Index ist der JIF durch die praktizierte Nutzung in Verruf geraten. Garfield selbst wies darauf hin, dass der JIF für das Ranking von Journalen entwickelt wurde und die Praxis der Bewertung der Leistung eines Autors mit Hilfe des JIF sehr problematisch ist (Garfield 2005). Auf den Websites von Thomson Reuters und Scopus findet man viele Informationen, wie die angebotenen Analysen zu verwenden sind. Auf mögliche Falschanwendungen wird hingewiesen und ausdrücklich vor Missbrauch gewarnt[25].

Neben der falschen Anwendung der Zitationsanalysen trägt die verwendete Datenbasis dazu bei, dass die errechneten Impact-Faktoren nicht immer ein realistisches Maß sein können. Obwohl die Zahl der ausgewerteten Zeitschriften sehr hoch erscheint, wird trotzdem nur ein mehr oder weniger geringer Teil der existierenden Zeitschriften einbezogen, deren Gesamtanzahl im Moment auf 26 000 (Vaas 2010) geschätzt wird.

Für den Science Citation Index werden bewusst die in den Fachdisziplinen renommiertesten ausgewählt, was sicherlich für den größten Teil der enthaltenen Zeitschriften gilt, aber trotzdem nicht in jedem Fall objektiv sein kann. Außerdem wird im Science Citation Index der Schwerpunkt bewusst auf Naturwissenschaften und Medizin gelegt. Die Mehrheit der Zeitschriften ist im Science Citation Index nicht enthalten. Scopus verwendet zwar eine wesentlich größere Datenbasis, kann aber mit 19 000 (die Angaben schwanken) auch nicht annähernd alle einbeziehen.

Zusammenfassend bleibt festzustellen, dass die Bewertung von wissenschaftlicher Leistung auf der Basis von Zitationsanalysen aus verschiedenen Gründen wie unvollständiger Datenbasis, falschen Schlussfolgerungen aus den Analysen, fehlenden Kontextinformationen und der falschen Annahme, dass Zitation immer eine Belohnung ist, zwar problematisch ist, sich aber trotzdem durchgesetzt hat. Richtig angewendet liefert die Zitationsanalyse und damit der Citation Impact zweifelsfrei wichtige Hinweise für die Bewertung wissenschaftlicher Leistungen.

1.2.4 Der Download Impact

Bei der Zitationsanalyse wird von inhaltlichen Bestandteilen von wissenschaftlichen Publikationen ausgegangen. Die Zitate sind in Form von Referenzlisten, Fußnoten oder ähnlichem fest im Text verankert und können, einmal veröffentlicht, nicht mehr verändert werden. Sind sie erst einmal identifiziert, werden sie gezählt und ausgewertet. Dabei ist es unerheblich, ob die Publikation digital oder auf Papier vorliegt.

[25] Siehe http://wokinfo.com/benefits/essays/, gelesen 30.3.2011.

So wie Zitationen als „Belohnung“ für eine Publikation interpretiert werden können, kann man auch davon ausgehen, dass die Nutzung einer Publikation ein Interesse daran ausdrückt. Daten über die Nutzung von Publikationen werden ständig neu generiert und sind nicht in anderen Publikationen enthalten. Nutzungsdaten entstehen bereits für Publikationen auf Papier, z. B. durch Verkauf oder durch Ausleihen in einer Bibliothek. Noch mehr ins Blickfeld von Untersuchungen gelangten sie im Zusammenhang mit elektronischen Publikationen, da deren Nutzung viel besser protokolliert werden kann. Darmoni et al. untersuchten das Verhalten der Leser einer digitalen Bibliothek mit Volltexten von 64 Journalen. Die Zugriffe darauf über ein Intranet wurden als „electronic consultations“ aufgzeichnet. Das Konzept der Zitation als Belohnung wurde auf einen Nutzungsvorgang übertragen und aus der Anzahl der Nutzungsvorgänge der Reading Factor für Journale gebildet (Darmoni et al. 2002). Ähnlich wurde bei der Bildung des Usage Impact Factors für Journale an der California State University vorgegangen (Bollen und van de Sompel 2008). Bei beiden Studien handelte es sich um Nutzungsdaten, die in abgeschlossenen elektronischen Netzen entstanden und erhoben wurden.

Wird auf eine Website im Internet mittels eines Browsers zugegriffen, wird der Inhalt der Website auf den Computer kopiert. Schon das wird als Download bezeichnet, egal, ob die Daten nur im Browser sichtbar sind oder anschließend lokal gespeichert werden. Alle Downloads, ob es nun Volltexte einer Publikation, Metadaten einer Publikation[26] oder beliebig andere Inhalte sind, werden protokolliert. Was dabei alles aufgezeichnet wird, ist im Abschnitt 2.1 zu lesen. Downloads von Publikationen zeigen wie die „electronic consultations“ eine Nutzung und damit ein Interesse an. Handelte es sich bei den Studien auch nicht um Downloads aus dem Internet, so wurden doch digitale Publikationen oder Teile davon kopiert und in diesem Sinne handelt es sich auch um Downloads. Impact-Maße, die auf der Analyse von Downloads basieren, werden als Download Impact zusammengefasst. Dazu gehört auch einfach die Anzahl der Downloads einer Publikation in einer bestimmten Zeiteinheit. Die Begriffe Zugriff und Download werden synonym verwendet. Downloadzahlen sind Nutzungsdaten einer Publikation.

Von diesem Ansatz gingen auch Brody et al. aus. Sie untersuchten die Downloads eines Open Access Repository, welches die Publikationen im Internet frei anbietet und wiesen die Korrelation des Download Impact in Form von Zugriffshäufigkeit im ersten Jahr nach der Publikation und späterem Citation Impact für arXiv.org[27], einem Repository hauptsächlich für Mathematik, Physik und Informatik, nach. Daraus folgerten sie, dass der Download Impact als Indikator für den Citation Impact angesehen werden kann (Brody et al. 2006). Es gilt laut Brody et al.: Je höher die Zugriffshäufigkeit, um so öfter wird das Dokument zitiert werden. Damit könnte die Zugriffshäufigkeit als Vorhersage für einen zu erwartenden Impact im Sinne von Zita-

[26] In dieser Arbeit wird der Begriff „Publikation“ bei einer elektronischen Publikation so verstanden, dass er die Gesamtheit aller Komponenten unfasst und so ein komplexes digitales Objekt bildet. Die Metadaten als einer Komponente davon enthalten beschreibende, technische und administrative Angaben zur Publikation (siehe auch (DINI 2010a), Glossar).

[27] http://archivx.org/, gelesen 28.3.2011.

tionshäufigkeit angesehen werden. Wie sich in späteren Untersuchungen zeigte, kann diese Behauptung nicht in solcher Allgemeinheit aufrechterhalten werden. Im Abschnitt 1.3.4 wird darauf ausführlicher eingegangen.

In den vorangegangenen Beispiele von Darmoni et al, Bollen und van de Sompel wurden Impact-Maße für Publikationen aus der Zählung von Downloads in Analogie zur Zählung von Zitationen gebildet. Der zweite Ansatz in der Zitationsanalyse, das Netzwerkmodell, kann auch unter bestimmten Annahmen auf elektronische Publikationen und deren Downloads übertragen werden. Beim Thema Citation Impact fiel bereits der Begriff PageRank-Algorithmus, der hier etwas näher erläutert werden soll. Um einen Bezug zur Praxis herzustellen, wird zuerst auf eine bekannte Anwendung im Internet eingegangen.

Mit der Verbreitung des Internets ging die Entwicklung von Suchmaschinen einher und es wurde für die Sichtbarkeit von Websites essentiell, dass sie in deren Index aufgenommen werden und an welcher Position sie im Suchergebnis erscheinen. Dazu müssen neben den technischen Möglichkeiten, das Internet abzusuchen und eine Dokumentenbasis aufzubauen, Algorithmen entwickelt werden, welche die Websites nach bestimmten Prinzipien bewerten, um sie anschließend in geordneter Reihenfolge als Suchergebnis päsentieren zu können.

Ein Bewertungsprinzip ist, dass von den Websites, die gleich gut auf den Suchbegriff passen, also relevant sind, die Website mit der höchsten Wahrscheinlichkeit, dass ein sich im Internet bewegender Nutzer gerade auf diese gelangt, den höchsten Rang erhält. Dieses Prinzip wurde von Larry Page und Sergej Brin 1998, den Gründern von Google[28], in Form des PageRank-Algorithmus, der bei der Google-Suche verwendet wird, umgesetzt (Brin und Page 1998)[29]. Sie griffen dabei auf eine Idee von Leo Katz zurück, der im Jahr 1953 einen neuen Index für den Status in einem sozialen Netzwerk kreierte (Katz 1953). Katz geht bei der Berechnung des Status davon aus, dass es wichtig ist, wieviel Kanten zu einem Knoten führen und von welchem Knoten diese Kanten ausgehen. Auf Websites wurde das Modell so übertragen, dass Websites die Knoten und Links von und zu diesen Websites die Kanten des Netzwerks bilden. Das Gewicht einer Website ist um so höher, je größer die Anzahl der Links ist, die auf sie zeigen. Ein Link wird umso höher bewichtet, je höher das Gewicht der Website ist, von der er kommt. Bei der Google-Suche stehen Websites allegmein im Fokus, im Internet vorhandene wissenschaftliche Publikationen werden wie alle anderen Websites in das Ranking einbezogen.

Geht man von folgenden Annahmen aus, können statt Websites und Links auch Publikationen im Internet und ihre Downloads mit einem Netzwerkmodell analysiert werden: Die erste Annahme ist, dass mit dem Download einer Publikation das Interesse an der Publikation bekundet wird. Als zweite Annahme wird davon ausgegangen, dass zwischen den Publikationen, die ein Nutzer in zeitlicher Nähe auf seinem Browser zur Kenntnis nimmt, eine Beziehung besteht. So entsteht ein Netzwerk mit digitalen Publikationen als Knoten

[28] Google: http://www.google.de, gelesen 30.3.2011.

[29] Welche Faktoren bei der Google-Suche die Relevanz bestimmen, ist nicht bekannt.

und den Beziehungen als Kanten. Wie bei Websites und Links ist es von Bedeutung, mit welcher anderen Publikation eine Beziehung besteht. Für diese Art von Netzwerken wurden die Begriffe „Usage Network“ oder „Reader Generated Network“ geprägt. Auf der Grundlage dieses Netzwerkmodells konnten Bollen et al. bei der Analyse der Nutzung einer großen digitalen Bibliothek[30] einen dem PageRank ähnlichen Algorithmus (Bollen et al. 2003) verwenden. Der Status einer Publikation wird hier wie beim PageRank dadurch bestimmt, zu wieviel anderen Publikationen Beziehungen bestehen und wie hoch deren Status ist. Es wurde ein Ranking von Journalen für zwei verschiedene Zeiträume durchgeführt, so dass neben dem Ranking auch ein Nutzungstrend erkennbar ist. Eine andere Anwendung dieses Netzwerkansatzes ist die Aufdeckung von Strukturen in der Wissenschaft in Form von Journal Maps oder Subject Categories, d. h. Gruppen von Zeitschriften oder Fachgebieten, die enge Beziehungen aufweisen (Bollen und van de Sompel 2008).

Jenseits aller Rankings ist es für Autoren immer wichtig zu wissen, ob ihre Publikationen zur Kenntnis genommen werden, was sich bei digitalen Artikeln im Download Impact widerspiegelt. Der Download Impact nimmt eine völlig andere Rolle als der Citation Impact ein, der nach wie vor die wirklich wichtige Zielgröße für Autoren ist. Im Gegensatz zum Citation Impact, der erst mit zeitlicher Verzögerung ermittel wird, kann der Download Impact eines Artikels jedoch ein Anzeichen für das Interesse an der Publikation sein,.

1.3 Open Access

Mit Open Access (OA) wird ein Prinzip des elektronischen wissenschaftlichen Publizierens bezeichnet, welches, alternativ zu konventionellen Publikationsformen, einen ungehinderten Zugang zu wissenschaftlichen Erkenntnissen gewährleistet. Publikationen, die unter Bedingungen publiziert werden, welche das OA-Prinzip einhalten, werden OA-Publikationen genannt. Das Prinzip, Zugang nur gegen eine Gebühr zu gewährleisten, wird als Toll Access (TA) bezeichnet.

Die Bewegung, die sich die Umsetzung des OA-Prinzips zum Ziel gesetzt hat, begann laut „Timeline of the Open Access Movement“[31] in den 1960er Jahren in den USA. Es wurde zwar schon bedeutend früher thematisiert, dass finanzielle Barrieren beim Zugang zu Publikationen den wissenschaftlichen Fortschritt behindern, aber erst mit dem Aufkommen von elektronischen Medien und der praktisch unbegrenzten Möglichkeit der Herstellung von Kopien entstanden die technischen Voraussetzungen, diese Barrieren zu überwinden. Im Jahr 1969 wurde damit begonnen, elektronische Publikationen auf FTP-Archiven und im Vorgänger des Internet ARPANET[32] für jeden, der über die technischen Möglichkeiten verfügte, frei zugänglich zu machen. Der kleine Kreis derer, die daran teilhaben konnten, vergrößerte sich erst mit der Entstehung und Verbreitung

[30] Es wurden Nutzungsdaten der Los Alamos National Laboratory's Research Library (LANL) analysiert.

[31] http://oad.simmons.edu/oadwiki/Timeline, gelesen 4.1.2010.

[32] ARPANET: Advanced Research Projects Agency Network.

des Internet in den 1990er Jahren[33]. In dieser Zeit wurden die ersten elektronischen Zeitschriften veröffentlicht und einer der heute noch wichtigsten Prototypen eines OA-Servers für Pro- und Preprints arXiv.org ging online. Im Jahr 1999 wurde die Open Archives Initiative[34] gegründet, die Standards für die Interoperabilität im Internet entwickelt. Ein wichtiger Motor der OA-Bewegung war und ist SPARC[35], eine internationale Allianz wissenschaftlicher Bibliotheken. Das Webangebot von SPARC ist eine Informationsplattform für alle an der OA-Bewegung beteiligten und interessierten Akteure.

Die wohl kürzeste und deshalb umgangssprachlich meist gebrauchte Erklärung, was OA bedeutet, lautet: Freier und kostenloser Zugang zu wissenschaftlichen Publikationen. Damit wird das OA-Prinzip aber nur sehr vereinfacht widergegeben. Eine umfassende Beschreibung ergibt sich aus den bisher drei wichtigsten Dokumenten, welche die Open Access Initiative im Verlauf ihrer Entwicklung hervorgebracht hat:

- "Budapest Open Access Initiative"[36], kurz Budapester Erklärung
- "Bethesda Statement on Open Access Publishing"[37], kurz Bethesda Statement
- "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities"[38], kurz Berliner Erklärung

Diese drei Dokumente gelten als Meilensteine der OA-Bewegung. Peter Suber, einer der Hauptinitiatoren, bezieht sich auf diese drei Dokumente in Form der „BBB definition of open access" (Budapest-Bethesda-Berlin) (Suber 2004). Er setzt sich mit der Kritik, es gebe keine Definition OA, auseinander und legt dar, dass sich die Inhalte der Deklarationen zwar in Teilaspekten unterscheiden, in den essentiellen Punkten aber übereinstimmen:

- "OA content must be free of charge for all users with an internet connection"
- "OA removes both price and permission barriers"

Peter Suber selbst relativiert den zweiten Punkt in der Form: "We should agree that OA removes some permission barriers." Die Diskussion einer Definition des Prinzips OA ist vielleicht noch nicht abgeschlossen, klar ist aber das Ziel, Wissen ohne Einschränkungen in digitaler Form zu verbreiten und zu erhalten.

---

[33] Tim Berners-Lee, der 1989 das Hypertext Transfer Protokoll http und 1990 den ersten Browser entwickelte, gilt als der Begründer des Internets.

[34] http://www.openarchives.org/, gelesen 28.3.2011.

[35] SPARC: Scholarly Publishing and Academic Resources Coalition, http://www.arl.org/sparc/index.shtml, gelesen 14.1.2011.

[36] http://www.soros.org/openaccess/, gelesen 13.1.2011.

[37] http://www.earlham.edu/~peters/fos/bethesda.htm, gelesen 13.1.2011.

[38] http://oa.mpg.de/lang/de/berlin-prozess/berliner-erklarung/, gelesen 13.1.2011.

Aus dem OA-Prinzip ergeben sich neben der Gewährleistung der Zugriffsbedingungen eine Reihe von technischen und rechtlichen Anforderungen, die im Bethesda Statement formuliert wurden. So muss garantiert sein, dass mindestens auf einem vertrauenswürdigen Repository eine vollständige, zur Langzeitarchivierung geeignete Kopie einer OA-Publikation inklusive zusätzlicher Materialien und einer Genehmigung zur OA-Publikation vorhanden sein muss. Des Weiteren müssen die Autoren den Nutzern umfassende unwiderrufliche Nutzungs- und Verteilungsrechte unter der Bedingung, dass die Urheberschaft genannt wird, einräumen. Bereits in der Budapester Erklärung werden zwei verschiedene Strategien des Open Access empfohlen, Self Archiving und OA Zeitschriften.

1.3.1 Self Archiving

In der Budapester Erklärung heißt es dazu:

"First, scholars need the tools and assistance to deposit their refereed journal articles in open electronic archives, a practice commonly called, self-archiving. When these archives conform to standards created by the Open Archives Initiative, then search engines and other tools can treat the separate archives as one. Users then need not know which archives exist or where they are located in order to find and make use of their contents."

Unter dem Begriff „Archiving" wird, da es im amerikanischen Kontext zu verstehen ist, das zur Verfügung stellen verstanden. Das Konzept des Self Archiving wurde durch die drei Möglichkeiten

- Individual Self Archiving,
- Institutional Self Archiving und
- Central Self Archiving

konkretisiert (Severiens und Hilf 2004). Diese Strategien des Self Archiving werden als „Green Road to Open Access" bezeichnet.

Unter Individual Self Archiving versteht man das Veröffentlichen von wissenschaftlichen Publikationen einzelner Autoren in Form persönlicher Websites oder Arbeitsgruppen auf Institutsservern, Arbeitsgruppenservern oder den Servern von Onlinediensten, ohne dass von einer zentralen Website aus die einzelnen Artikel verlinkt sein müssen. Meist sind diese sehr begrenzten Archive für die Zusammenarbeit von kleinen Gruppen gedacht. Von „fremden" Nutzern werden diese Archive meist nur durch Suchmaschinen, die unspezifisch in allen Websites suchen, aufgefunden, da sie nicht über Schnittstellen verfügen, die die Suche über Wissenschaftsportale ermöglichen. Die meisten Formen des Individual Self Archiving gewähren zwar einen ungehinderten Zugriff, können aber die technischen Anforderungen, die im Bethesda Statement genannt wurden, nicht erfüllen.

Institutional Self Archiving ist die Veröffentlichung auf Servern von Institutionen sehr unterschiedlicher Größe. Sie bieten für Nutzer und Autoren wesentlich mehr Komfort wie zentrale Suchportale, Langzeitarchivierung, Autoren- und Herausgebersupport und besitzen oft die empfohlene OAI-Schnittstelle, die das Zugreifen von übergeordneten, zum Teil weltweit agierenden Suchdiensten ermöglichen. Als Beispiel sei hier

die OAIster Database[39] genannt, deren Metadaten über das OAI-PMH-Protokoll[40] gesammelt werden. Der größte Teil der Metadaten stammt aus Bibliothekskatalogen, aber auch beliebige andere Institutionen können ihre Daten zur Verfügung stellen. Institutional Self Archiving findet auf sogenannten Institutional Repositories statt, welche im Abschnitt 1.4 genauer behandelt werden. Institutional Repositories sind mit ihrem höheren Standard auch für einzelne Autoren und kleinere Gruppen besser für das Individuell Self Archiving geeignet als die meisten individuellen Möglichkeiten.

Das früheste, bekannteste und mittlerweile auch am meisten untersuchte Beispiel für Central Self Archiving ist der bereits erwähnte Server arXiv. Schon bevor das Internet aufkam, etablierte Paul Ginsparg im Jahr 1991 ein Repository für Preprints aus dem Bereich der Physik, welches bald um weitere Fächer erweitert wurde. Es wurde sehr schnell ein wichtiger Bestandteil der wissenschaftlichen Kommunikation der beteiligten Communities und erlaubte die Veröffentlichung von Artikeln, lange Zeit bevor diese von Zeitschriften akzeptiert und veröffentlicht wurden. Mittlerweile gibt es eine große Anzahl fachspezifischer Repositories, die Central Self Archiving anbieten und den technischen Standards des Bethesda Statements entsprechen.

Da sich Institutional Self Archiving und Central Self Archiving, was die technischen Standards und die Möglichkeiten der Interoperabilität anbetrifft, inzwischen, von Ausnahmen abgesehen, kaum noch unterscheiden und Institutional Repositories eine immer größere Verbreitung finden, wird die Unterscheidung ihre praktische Bedeutung mehr und mehr verlieren. Für wissenschaftlichen Suchportale macht es technisch keinen Unterschied, von welcher der beiden Archiving-Formen Publikationen stammen (Harnad 2006).

Von Verlagen wird Self Archiving von Artikeln, die bei ihnen erschienen sind, verständlicherweise kritisch betrachtet und sehr differenziert behandelt. Die Regelungen reichen von striktem Verbot bis zur kompletten Freigabe für OA. Dazwischen existieren unzählige Varianten und Bedingungen. Was Verlage den Autoren gestatten, wird weitgehend in der sogenannten SHERPA/RoMEO[41]-Liste erfasst. Durch die Auskünfte aus der Liste sollen Autoren eine gewisse Sicherheit beim Self Archiving erhalten und zur OA-Publikation ermutigt werden. Es gibt auch die Möglichkeit, dass Autoren gegen eine Gebühr die Erlaubnis vom Verlag erhalten, sofort oder nach einem Embargozeitraum auf einem OA-Repository zu publizieren. Eine Liste dieser Verlage findet man auch bei SHERPA/RoMEO[42].

[39] http://www.oclc.org/oaister/, gelesen 13.1.2011.

[40] OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting.

[41] SHERPA war ein Projekt britischer Bibliotheken und Universitäten mit dem Ziel der Etablierung von OA-Repositories, http://www.sherpa.ac.uk/projects/sherpa.html. RoMEO ist eine von vielen Fortsetzungen davon. Es gibt eine englische (http://www.sherpa.ac.uk/romeo/) und eine deutsche (http://www.dini.de/wiss-publizieren/sherparomeo/) Liste, gelesen 14.1.2011.

[42] http://www.sherpa.ac.uk/romeo/PaidOA.html, gelesen 14.1.2011.

1.3.2 Open Access-Zeitschriften

In der Budapester Erklärung heißt es weiter:

“Second, scholars need the means to launch a new generation of journals committed to open access, and to help existing journals that elect to make the transition to open access. Because journal articles should be disseminated as widely as possible, these new journals will no longer invoke copyright to restrict access to and use of the material they publish. Instead they will use copyright and other tools to ensure permanent open access to all the articles they publish ...”

OA-Zeitschriften sind eine neue Generation wissenschaftlicher Zeitschriften, auf die der Zugriff über das Internet erfolgt, ohne Einschränkungen permanent möglich und für Leser kostenfrei ist. Diese Strategie wird als „Golden Road to Open Access“ bezeichnet. Im Gegensatz dazu bieten die meisten traditionellen Verlage ihre Zeitschriften gegen Subskriptionsgebühren als TA-Journale an.

Der überwiegende Teil wurde von Beginn an als OA-Zeitschrift konzipiert, andere existierten bereits als traditionelle TA-Zeitschriften und wurden in OA-Zeitschriften umgewandelt. Eine neue Entwicklung ist die Gründung von OA-Zeitschriften durch traditionelle renommierte, bisher ausschließlich im TA-Bereich agierende Verlage. Ob diese neuen OA-Zeitschrifte vom Prestige der Verlage profitieren können und z.B. in den Science Citation Index aufgenommen werden, wie in Aussicht gestellt wird, bleibt abzuwarten.

Seit der Budapester Erklärung sind Tausende von OA-Zeitschriften entstanden. Das Directory of Open Access Journals[43] erfasst seit 2004 wissenschaftliche OA-Zeitschriften, die ausschließlich nach dem OA-Prinzip arbeiten und über eine Qualitätskontrolle durch Herausgeber oder ein Peer-Review-System verfügen. Herausgeber können sich für die Aufnahme bewerben und Leser können Zeitschriften vorschlagen. Im Mai 2010 waren es bereits 5000. Der überwiegende Teil verfügt über ein Begutachtungssystem, wobei meist das bei traditionellen Verlagen anerkannte Peer Review verwendet wird (Müller 2009).

Es gibt eine Reihe von OA-Zeitschriften, die Eingang in den Science Citation Index gefunden haben und überraschend hohe JIF erhielten. Man muss aber davon ausgehen, dass bisher nur ein geringer Teil der OA-Zeitschriften in die Berechnung des JIF einbezogen wird, wobei berücksichtigt werden muss, dass ein JIF frühestens im 3. Erscheinungsjahr vergeben werden kann. Im WoS gibt es nur einen allgemeinen Hinweis darauf, aber keine Zahlen. Das Directory of Open Access Journals enthält keine Angaben darüber, ob eine OA-Zeitschrift dort gelistet wird.

Zunehmend mehr traditionelle Verlage, die bisher keinen freien Zugriff auf ihre Publikationen gestatteten, erkennen inzwischen, dass sie sich der OA-Bewegung nicht mehr völlig verschließen können. Der Grund dafür ist nicht, dass Verlage OA als den besseren Weg erkannt haben, sondern dass bereits viele wissenschaft-

[43] http://www.doaj.org/, gelesen 13.1.2011.

liche Institutionen ihre Autoren zu OA verpflichten oder Anreize dafür geben. Viele Verlage bieten ihren Autoren an, akzeptierte Artikel gegen eine Gebühr für OA freizugeben. Sie zählen dadurch aber nicht zu OA-Verlagen.

1.3.3 Akzeptanz von Open Access

Als die ersten Ideen zu OA entstanden, ging es darum, bereits publizierte Literatur, die auf Papier existierte, allgemein zugänglich zu machen. Seit der Existenz des Internets wird das OA-Prinzip ausschließlich auf digitale Publikationen angewendet. Die frühen Ideen von der Verfügbarkeit des Wissens für alle werden durch Self Archiving von Pre- und Postprints und OA-Zeitschriften verwirklicht. Die Schnelligkeit des Mediums machte besonders das Publizieren von Preprints attraktiv.

Da der Zeitraum von der Einreichung eines Artikels bis zu dessen Publikation in der Zeitschrift oft sehr groß ist, aber eine schnelle Publikation von Forschungsergebnissen angestrebt wird, wurde die technische Möglichkeit, elektronische Artikel sofort als Preprint online zur Verfügung zu stellen, sehr schnell populär. Diese Preprints haben noch kein Peer Review durchlaufen und der Leser kann sich nicht darauf verlassen, dass der Artikel einer Begutachtung standhält, wird doch ein großer Teil der bei einer Zeitschrift eingereichten Artikel von den Herausgebern abgelehnt. Leser fühlen sich verunsichert, was die Qualität der Artikel anbetrifft, und können sich häufig nur an Namen und Institutionen bei der Einschätzung orientieren. Ein weiterer Grund, OA-Publikationen kritisch zu betrachten, liegt darin, dass vermeintliche Preprints in Wirklichkeit der sogenannten grauen Literatur zuzurechnen und nicht dazu gedacht sind, durch einen Verlag veröffentlicht zu werden. Oft können sie von echten Preprints nicht unterschieden werden. Autoren befürchten einen Verlust an Ansehen, wenn sie ihre Artikel in einer Reihe mit vielleicht minderwertigen Artikeln publizieren. Ausserdem sehen sie die OA-Publikation mit Aufwand verbunden und haben Vorbehalte hinsichtlich Qualität der Archivierung und der Wahrung ihrer Urheberschaft.

Leider werden diese Vorbehalte gegenüber dem Self Archiving von Rezipienten und Autoren auch auf OA-Zeitschriften übertragen, obwohl die angeführten Argumente hier keine Gültigkeit besitzen. Trotz der Anerkennung von OA als wichtiges und richtiges Prinzip ist eine ambivalente Haltung dazu verbreitet. Untersuchungen über die Einstellung von Wissenschaftlern zu OA kamen zu dem Ergebnis, dass eine breite Lücke zwischen der bekundeten Absicht, nach dem OA-Prinzip zu publizieren, und dem tatsächlichen Verhalten besteht (Hess et al. 2007; Swan 2006a). Jeweils eine Mehrheit Befragter einer 2006 durchgeführten Studie befürchtet eine Gefährdung von Stellung und Beförderung und Verringerung von Chancen auf Forschungsmittel. Ebenfalls eine Mehrheit beklagt, dass nur wenige OA-Zeitschriften im Science Citation Index enthalten sind und führt das als Grund gegen OA auf. Andererseits gibt auch eine Mehrheit an, sehr wohl auf OA-Artikel zuzugreifen (Mann et al. 2009).

Obwohl seit der Budapester Erklärung Tausende von OA-Zeitschriften entstanden, ist deren Reputation bis auf Ausnahmen bisher gering geblieben. Der Zeitfaktor spielt zwar eine Rolle, da sie sich Reputation erst verdienen müssen, kann aber das Akzeptanzproblem nicht ausgleichen. Um das Akzeptanzverhalten zu ver-

ändern, sind nicht nur gezielte Fördermaßnahmen, sondern ein generelles Umdenken im gesamten Wissenschaftsbetrieb erforderlich. Das gilt zwar allgemein für OA, für OA-Zeitschriften aber im Besonderen[44].

Wie schon anklang, ist die Diskussion über Definition, Strategien und Möglichkeiten von OA noch längst nicht abgeschlossen. Der nächste Schritt, nämlich die Umsetzung in die Praxis und damit die Ablösung herkömmlicher Publikationsarten befindet sich noch immer im Anfangsstadium und die Widerstände auf der Seite von Verlagen, die mit wissenschaftlichen Publikationen Geld verdienen, aber auch die Zurückhaltung auf der Seite von Autoren sind beträchtlich.

Trotz allem kann konstatiert werden, dass OA inzwischen ein fester Bestandteil des weltweiten Wissenschaftsbetriebes geworden ist. Davon zeugen die Vielzahl von Initiativen und Deklarationen, von denen viele in der schon erwähnten und ständig fortgeführten Timeline aufgeführt sind. Darunter befinden sich bei weitem nicht nur wissenschaftliche Institutionen, sondern auch internationale Organisationen wie UNESCO[45] und OECD[46]. Wie erfolreich die Bemühungen sind, kann daran gemessen werden, wie groß der Anteil von OA-Publikationen an der Gesamtzahl von Publikationen ist. Hier kann nur geschätzt werden. Dazu wurden über die Jahre eine Reihe von Studien mit unterschiedlichen Methoden durchgeführt. Björk et al. zogen eine zufällige Stichprobe von Artikeln, die 2008 in Zeitschriften mit Peer Review publiziert wurden und kamen zu dem Ergebnis, dass 2009 über alle Disziplinen hinweg 20,4% der Artikel als OA verfügbar waren (Björk et al. 2010).

1.3.4 Open Access und Impact

Wie aus den vorangegangenen Abschnitten bereits ersichtlich, sind die Chancen für OA-Zeitschriften, einen hohen JIF zu erlangen, bisher gering. Sind Autoren aus Gründen der Reputation darauf angewiesen, in Zeitschriften mit hohem JIF zu veröffentlichen, müssen sie das nach wie vor in der Regel in traditionellen TA-Zeitschriften tun. Dass es trotzdem im Hinblick auf den Citation Impact von Vorteil ist, neben der Veröffentlichung unter TA-Bedingung auf dem Weg des Self Archiving Post- oder Preprints frei zur Verfügung zu stellen, zeigt eine Reihe von Studien.

Die ersten Studien, in denen Artikel aus Conference Proceedings der Jahre 1989 bis 1999 analysiert wurden, erschienen 2001 (Lawrence 2001). Hier wurde festgestellt, dass die Anzahl der Zitationen von online verfügbaren Artikeln bis zu 336 % derer, die nicht online waren, betrug. Online bedeutete hier nicht unbedingt OA.

Darauf und auf weitere Studien beziehen sich Harnad et al. in ihrem vielbeachteten Artikel aus dem Jahr 2004, in dessen Abstract behauptet wird: „OA articles have significantly higher citation impact than non-OA

[44] Umfassende Informationen zur Akzeptanz von OA Journalen findet man in (Weishaupt 2009).

[45] Ein Handbuch zu OA wurde von der Deutschen UNESCO-Komission herausgegeben (UNESCO 2007) und im gleichen Jahr die Kronberg-Deklaration veröffentlicht: http://www.unesco.de/kronberg_declaration.html, gelesen 13.1.2011.

[46] http://www.oecd.org/site/0,3407,en_21571361_38415463_1_1_1_1_1,00.html, gelesen 4.1.2011.

articles.“ (Harnad et al. 2004). Vorausgegangen waren, außer denen von Lawrence, unter anderem Studien von Kurtz et al und Odlyzko (Kurtz et al. 2005a; Kurtz et al. 2005b; Odlyzko 2002), in denen ähnliche Effekte nachgewiesen wurden. In der Folge wurde vom „Open Access Citation Advantage“ gesprochen.

Insgesamt wurden viele Studien an unterschiedlichen Datenbeständen durchgeführt, deren Ergebnisse den „OA Citation Advantage“ bestätigten. Es gab aber auch Studien, die ihn in Frage stellten und die positiven Ergebnisse auf methodische Fehler zurückführten (Craig et al. 2007). Selbst Kurtz und Henneken revidierten frühere Aussagen (Kurtz und Henneken 2007). Wie auch Craig et al. nennen sie drei mögliche Gründe für die beobachteten Effekte:

a) Ein früherer Zugriff führt dazu, dass der Zeitraum für Zitation vergrößert ist (Early Access oder Early View)[47].

b) Es besteht die Tendenz, dass Autoren vor allem ihre besten Artikel OA publizieren und vor allem erfolgreiche Autoren ihre Artikel OA publizieren (Quality oder Selection Bias).

c) OA-Artikel werden mehr gelesen und deshalb mehr zitiert

Ihre Untersuchungen zeigten, dass nicht c), sondern a) und b) einen höheren Citation Impact verursachten. In einer von Davis et al. durchgeführten Studie wurden a) und b) ausgeschlossen, in dem in einer TA-Zeitschrift eine zufällige Auswahl von Artikeln den OA-Status erhielten. Bei diesen Artikeln war die Anzahl der Downloads wesentlich höher, aber nicht die Anzahl der Zitationen (Davis et al. 2008).

Alma Swan, eine Protagonistin der OA-Bewegung, veröffentlichte eine Zusammenfassung von 31 Studien zum „Open Access Citation Advantage“, von denen 27 dafür sprechen und 4 keinen Effekt finden können (Swan 2010). Ohne es explizit zu formulieren, wird durch das Zahlenverhältnis nahegelegt, dass die Hypothese „OA führt zu einem höherem Citation Impact“ zutrifft. Es handelt sich hier aber um Studien mit unterschiedlichen Kontexten und Methoden, so dass eine solche Schlussfolgerung nicht gezogen werden kann.

Die Studien pro und contra „Open Access Citation Advantage“ lösen sich während der vergangenen Jahre gegenseitig ab und ein Ende der Diskussion ist nicht abzusehen[48]. Dass Artikel mit OA–Status wesentlich mehr Zugriffe und damit einen größeren Download Impact als Artikel mit TA-Status haben, was das Hauptmotiv für OA ist, konnte in den Studien zweifelsfrei nachgewiesen werden.

[47]Eine genaue Definition des Zeitfensters ist für bibliometrische Untersuchungen erforderlich.

[48] Eine Bibliographie zum Einfluss von OA und Downloads auf den Citation Impact findet man beim Open Citation Project http://opcit.eprints.org/oacitation-biblio.html#most-recent, gelesen 18.1.2011.

1.4 Institutional Repositories

1.4.1 Aufgaben

Die überwiegende Anzahl von Forschungsinstitutionen besitzt zum gegenwärtigen Zeitpunkt Dokumentenserver für die Sammlung von Forschungsergebnissen in digitaler Form und nutzt zu deren Verbreitung das Internet. Gebräuchlich ist hierfür der Begriff **„**Institutional Repository**“** (IR), der folgendermaßen definiert werden kann: „An institutional repository (IR) is an electronic system that captures, preserves, and provides access to the digital work products of a community.” (Foster und Gibbons 2005). Oder etwas weitergehend die Definition von SPARC[49], die den Zugriffsaspekt betont: „… an institutional repository is a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution, with few if any barriers to access.” (Crow 2002). Da der Begriff in Fachkreisen meist in der englischen Form benutzt wird, soll dieser auch hier verwendet werden.

Neben den IRs gibt es die sogenannten Disciplinary Repositories, die die digitalen Publikationen bestimmter Fachgebiete sammeln und diese ebenfalls ohne Zugriffsbeschränkungen anbieten. Disciplinary Repositories dienen dem Central Self Archiving. Institutional und Disciplinary Repositories, die einem bestimmten Qualitätsstandard genügen, sind im Directory of Open Access Repositories OpenDOAR[50] und im Registry of Open Access Repositories ROAR[51] (Anzahl im Januar 2011: 1800 bzw. 2000) verzeichnet.

Heute stellen IR oft den Kern des Dokumenten- und Publikationsservices einer Forschungseinrichtung dar. Als Serviceeinrichtung für eine Institution werden sie für die Veröffentlichung vieler verschiedener Arten von Publikationen genutzt. Die IR von Hochschulen sind häufig Weiterentwicklungen von Dokumentenservern, die vorrangig der Erstpublikation von Hochschulschriften wie Dissertationen, Habilitationsschriften und Abschlussarbeiten dienten. Deshalb bilden solche Hochschulschriften nach wie vor einen großen Anteil an den Publikationen von universitären IR.

Im Zuge der OA-Bewegung erlangten IR einen enormen Bedeutungszuwachs, da sie den Wissenschaftlern als Instrument des Institutional Self Archiving dienen und so Bestandteil der „Green Road to Open Access“ geworden sind. Die Nutzung von IR für das Institutional Self Archiving entwickelt sich positiv und der Anteil von Pre- und Postprints von Artikeln, die an anderer Stelle veröffentlicht werden sollen oder bereits veröffentlicht wurden, ist im Ansteigen begriffen.

Von Herausgebern von Zeitschriften und Serien wurden IR häufig neben der Papierform für die elektronische Zweitpublikation genutzt. Mehr und mehr wird jedoch die Papierform völlig durch die ausschließlich elek-

[49] SPARC: The Scholarly Publishing &Academic Resource Coalition, http://www.arl.org/sparc, gelesen 18.1.2011.

[50] Centre for Research Communications, University of Nottingham: http://www.opendoar.org/, gelesen 18.1.2011.

[51] University of Southhampton: http://roar.eprints.org/, gelesen 18.1.2011.

tronischen Publikation ersetzt. Ein oft gewählter Weg ist der Übergang von der Papierform zur elektronischen Form durch Digitalisierung alter Ausgaben und Weiterführung als Onlinepublikation. Neue Zeitschriften, Sammelwerke und Serien schließlich werden von vornherein als ausschließlich elektronische Publikation konzipiert, wobei die Möglichkeit des individuellen Drucks eine regelmäßige Anforderung der Herausgeber ist. Zum Aufgabenspektrum von IR zählt inzwischen ganz selbstverständlich die Publikation von allen Veröffentlichungen im Zusammenhang mit Konferenzen und Workshops. Alle diese Formen der Publikation sind der „Golden Road to Open Access“ zuzuordnen.

Zu diesen bisher aufgeführten Aufgaben kommen häufig noch andere, sehr individuell von der Institution abhängige Aufgaben hinzu. So ist die Bereitstellung von Materialien, die von Arbeitsgruppen genutzt werden, üblich. Auch sind nicht alle Publikationen von IR frei, sondern nur für bestimmte Mitarbeiterkreise zugänglich. Die Vielfalt an Aufgaben erzeugt höhere konzeptionelle und technologische Anforderungen an die Betreiber, als ursprünglich vorhersehbar war. Diesen Anforderungen kann ein IR nur dann genügen, wenn sein Aufbau gründlich vorbereitet und geplant wird. Software muss angepasst oder erst entwickelt werden. Die Modellierung und Implementierung eines speziell für die jeweilige Institution zugeschnittenen Workflows erfordert erheblichen Aufwand. Existierende Standards, z. B. die von DINI[52] regelmäßig aktualisierten (DINI 2010a), sind zu beachten. Die Empfehlungen von SPARC (SPARC 2002) in Form einer checklist sind nach wie vor aktuell[53].

Auch nach der Überführung des IR in den Routinebetrieb stehen Pflege und Weiterentwicklung der IT-Komponenten an. Neben Sammlung, Veröffentlichung und Verbreitung liegen auch Wahrung der Authentizität und Integrität und die Langzeitarchivierung der Publikationen in der Verantwortung des Betreibers (Schirmbacher 2005). Die Akquisition von Publikationen, Betreuung von Autoren und Herausgebern und nicht zuletzt die fachgerechte Eingabe der Metadaten sind regelmäßig anfallende Tätigkeiten. Damit diese Aufgaben bewältigen werden können, muss ein Team von Fachleuten aus Informatik und Bibliothekswesen ständig bereitstehen.

Obwohl schon in der Budapester Erklärung auf die Schaffung von Interoperabilität durch OAI-Schnittstellen hingewiesen wurde, erfolgt der Betrieb der einzelnen Server bisher weitgehend unabhängig voneinander. Durch die oft vorhandene OAI-Schnittstelle ist zwar eine Suche nach Metadaten in wissenschaftlichen Portalen wie das bereits erwähnte OAIster oder BASE[54] möglich, Mehrwertdienste wie Dublettenkontrolle oder Nutzungsstatistiken sind nicht vorhanden. Die Recherche nach Publikationen in IR gestaltete sich deshalb unkomfortabel, was der Akzeptanz bei Lesern nicht förderlich ist. Viel besser wäre ein gemeinsames Portal möglichst vieler IR. Die Potentiale der IRs zu bündeln und eine gemeinsame Open-Access-Infrastruktur

[52] DINI: Deutsche Initiative für NetzwerkInformation e.V., http://www.dini.de, gelesen 28.3.2011.

[53] Eine systematische Anleitung zum Aufbau eines IR findet man in (Dobratz und Müller 2009).

[54] BASE: Bielefeld Academic Search Engine, http://base.ub.uni-bielefeld.de/de/index.php, gelesen 28.3.2011.

aufzubauen ist das Ziel nationaler[55] und internationaler Projekte[56]. Als Ergebnis dieser Projekte wird ein wesentlicher Synergieeffekt für die OA-Bewegung erwartet, in dem man dem Ziel aus der Budapester Erklärung, möglichst viele Repositories als ein durchsuchbares Archiv anzusehen, entscheidend näher kommt.

1.4.2 Sichtbarkeit und Ranking im Internet

Was versteht man unter der Sichtbarkeit einer Website im Internet? Drèze und Zufryden definieren sie folgendermaßen: “Visibility is defined as the extent to which a user is likely to come across a reference to a company’s Web site in his or her online or offline environment.” (Drèze und Zufryden 2004). Verallgemeinert man diese Definition und bezieht sie nur auf das Internet, erhält man eine Definition, die in Übereinstimmung mit dem allgemeinen Sprachgebrauch steht: Sichtbarkeit einer Website im Internet ist die Wahrscheinlichkeit dafür, dass ein Nutzer zu dieser Website gelangt. In diesem Sinne soll hier der Begriff Sichtbarkeit verwendet werden.

Mit der Verbreitung des Internets entstand gleichzeitig eine Forschungsdisziplin, die die Messung von Webinhalten zum Gegenstand hat, die Webometrie. In der Webometrie werden häufig aus der Bibliometrie bekannte Methoden auf Webinhalte und ihre Beziehungen untereinander übertragen, wobei in Analogie zur Zitationsanalyse eine Analyse von Links stattfindet. Um diese Analogie von Zitation und Link sprachlich zu betonen, wurde der Begriff „Sitation“ als Synonym für „Link“ verwendet. Ingwersen übertrug das Konzept des Citation Impact auf Webinhalte und führte den Web Impact Factor (WIF) als Maß für die Sichtbarkeit von Websites[57] (Ingwersen 1998) ein. Der WIF wird durch das Verhältnis der Anzahl von Webpages, auf die es einen Link gibt, zur Gesamtzahl der Webpages der Website definiert. Dabei kann die Website z.B. ein IR sein und die dazugehörenden Webpages die Publikationen. Der WIF wird aber auch für ganze Institutionen und sogar für Länder berechnet. Da die Berechnung des WIF auf Daten beruht, die durch Suchmaschinen ermittelt wurden und nicht überprüfbar sind, gilt der WIF nicht als verlässliches Maß (Noruzi 2006).

Die Annahme, Links würden die Wahrscheinlichkeit, dass ein Nutzer zu einer Website gelangt, entscheidend beeinflussen, liegt auch dem bereits im Zusammenhang mit dem Download-Impact erwähnten PageRank-Algorithmus zugrunde. Von allen auf die Suchanfrage passenden Websites erhält die den höchsten Rang, die in einem Netzwerk aus Links und Websites den höchsten Status hat. Dieser Website wird die größte Relevanz zugeschrieben. Damit ist die Position einer Website im Ergebnis einer Suche, deren Algorithmus auf dieser Definition beruht, bereits ein Maß für Sichtbarkeit im Internet. Die Position im Suchergebnis wird bei den verschiedenen Suchmaschinen durch weitere Faktoren beeinflusst, z. B. die geografische Relevanz oder die Berücksichtigung der Suchgeschichte des Nutzers. Solche Abhängigkeiten werden nicht offengelegt, um Manipulationen zu erschweren. Die Bedeutung, die dem Ergebnis von Suchmaschinen für die Sichtbarkeit

[55] OA-Netzwerk: http://www.dini.de/projekte/oa-netzwerk/, gelesen 28.3.2011.

[56] DRIVER: http://www.driver-repository.eu/, gelesen 28.3.2011.

[57] In Verbindung mit dem Web Impact ist eine Website die Sammlung zusammenhängender Webpages.

zugemessen wird, wird schon durch den Begriff „Search Visibility“ deutlich, der Sichtbarkeit mit einem Suchergebnis gleichsetzt. Die gezielte Erhöhung der Sichtbarkeit geschieht mit dem Ziel, die Nutzungshäufigkeit zu erhöhen. Erwartungsgemäß gibt es eine hohe Korrelation von Sichtbarkeit in Form von Ranking im Suchergebnis und Nutzung[58].

Sichtbarkeit mit dem Ergebnis einer Suche gleichzusetzen, ist allerdings eine starke Vergröberung, da das Verhalten der Nutzer, also wie ein Nutzer im Internet navigiert, ebenfalls Einfluss hat. Dieses Verhalten wird nicht nur durch Links beeinflusst, sondern von vielen anderen Faktoren. Dazu gibt es Untersuchungen, die hauptsächlich den kommerziellen Bereich des Internets zum Gegenstand haben. So definieren Drèze und Zufryden einen „Internet Visibility Index“, der unter anderem von Werbung, Suchmaschinenergebnis, Links von anderen Seiten und Listung in einem Web-Verzeichnis abhängt (Drèze und Zufryden 2004). Betrachtet man Betreiber von IR, die wissenschaftliche Publikationen im Internet verbreiten, als Produzenten und die Rezipienten der Publikationen als Kunden, können die Ergebnisse vom kommerziellen Bereich zum Teil übertragen werden. So kann die Einteilung in psychologische und physische Faktoren, die die Sichtbarkeit im Internet beeinflussen, durchaus für die Sichtbarkeit wissenschaftlicher Publikationen nachvollzogen werden. Psychologische Faktoren sind die Verhaltensweisen der Nutzer im Umgang mit dem Internet, z. B. bei der Verwendung von Suchmaschinen und deren Resultaten. Zu den physischen Faktoren gehören Links, die Position im Ergebnis von Suchmaschinen und die Auflistung von Websites in Online-Verzeichnissen (Schmidt-Mänz und Gaul 2004).

Suchen Wissenschaftler nach Publikationen im Internet, nutzen sie häufig zuerst die nicht auf wissenschaftliche Inhalte spezialisierten Suchmaschinen. Die Recherche nach Autoren oder Titeln ist damit möglich, die nach bestimmten Inhalten ist dagegen mühselig. Der prozentuale Anteil von Suchergebnissen mit wissenschaftlichem Inhalt liegt im Promille-Bereich und es ist sehr unwahrscheinlich, diese in oberen Positionen des Rankings zu finden, was dazu führt, dass sie so gut wie nie wahrgenommen werden (Mayr 2009).

Gezielte inhaltliche Recherchemöglichkeiten bieten spezialisierte Suchen wie z. B. die des Web of Knowledge. Leider sind die Publikationen von IR kaum in solchen Suchmaschinen indexiert, was die Sichtbarkeit von OA-Publikationen drastisch verschlechtert. Die speziell für OA etablierten Suchdienste wie BASE oder OAIster sind Wissenschaftlern häufig unbekannt. Die Recherche nach Inhalten direkt auf OA-Repositories bietet im Allgemeinen nur geringe Aussicht auf Erfolg, da der Umfang an Publikationen eines Repositories begrenzt ist. Ein Portal, welches die OAI-Schnittstelle nutzt und die Publikationen möglichst vieler OA-Repositories von zentraler Stelle aus in gut erschlossener Form anbietet, könnte diese Situation verbessern[59]. Für die Sichtbarkeit der einzelnen Publikation ist hier wieder die Position im Suchergebnis wichtig.

[58] Siehe z. B. http://www.lightseeker.de/seo-erfolgsmessung-durch-rankingtraffic-korrelation/, gelesen 28.3.2011.

[59] In Deutschland ist die Schaffung eines solchen Portals im Projekt OA-Netzwerk geplant, siehe dazu auch (Malitz 2009).

Bei der Messung der Sichtbarkeit von Webservern wie Institutional oder Disciplinary Repositories gibt es neben dem WIF eine auf den Gegenstand spezialisierte Herangehensweise. Im "Ranking Web of World Repositories"[60], einer Initiative, die sich mit der quantitativen Analyse des Internets im Bereich der Wissenschaft beschäftigt und ein Ranking von Repositories durchführt, ist die Sichtbarkeit im Sinne der Anzahl der externen Links der wichtigste von vier Faktoren. Weitere Faktoren sind die Anzahl der Textfiles im PDF-Format, die Anzahl der von großen Suchmaschinen indexierten zugehörigen Webpages und die Aktualität der Publikationen auf der Basis von Google Scholar. Sichtbarkeit in diesem komplexeren Sinn wird hier als Qualitätskriterium für das Ranking verwendet.

Fasst man diese Aspekte der Sichtbarkeit, so wie sie oben definiert wurde, und des Rankings im Internet zusammen, kommt man für IR und deren Publikationen zu folgendem Ergebnis: Sichtbarkeit kann für einzelne Publikationen, Websites und ganze IR gemessen werden und Rankings, wie sie von Suchmaschinen durchgeführt werden, sind Ausdruck von Sichtbarkeit und beeinflussen die Nutzung. Sichtbarkeit kann durch Nutzung belegt werden, da es ohne Sichtbarkeit keine Nutzung gibt. Nutzungshäufigkeit und Sichtbarkeit sind miteinander korreliert, so dass im Einzelfall die Nutzungshäufigkeit einer Publikation trotz schlechter Sichtbarkeit hoch sein kann, in der Regel aber die Nutzungshäufikeit ein Indiz für die Sichtbarkeit ist.

1.4.3 Qualität von Institutional Repositories

Die Qualität eines IR muss unter unterschiedlichen Aspekten beurteilt werden. Der erste Aspekt betrifft die wissenschaftliche Qualität der Publikationen, die durch Betreiber weder eingeschätzt werden kann noch können sie darauf Einfluss nehmen. Durch die Bereitstellung von Begutachtungssystemen wird jedoch eine Unterstützung der Herausgeber bei der Qualitätssicherung geboten.

Der zweite Aspekt betrifft den Bestand an Publikationen. Dieser Bestand ergibt sich zuerst aus dem Bedarf von Autoren und Herausgebern der Institution an Publikationsmöglichkeiten, der ebenfalls von den Betreibern unabhängig ist. Der Bestand wird jedoch auch davon beeinflusst, inwieweit Betreiber auf Anforderungen, Vorstellungen und Wünsche von Autoren und Herausgebern eingehen können und wollen. In welchem Maß das IR durch die Publikationen den wissenschaftlichen Output der Institution widerspiegelt, hängt auch zuerst von Autoren und Herausgebern ab, kann aber durch gezielte Akquisition von Publikationen unterrepräsentierter Fachgebiete oder Struktureinrichtungen beeinflusst werden. Alle Maßnahmen, die die Steigerung der Akzeptanz zum Ziel haben, wie

- Flexibilität bei der Gestaltung des Online-Auftritts,
- Betreuung von Autoren und Herausgebern durch Unterstützung bei der technischen Erstellung der elektronischen Publikation,
- Bereitstellung von Informationen zu rechtlichen Fragen der Veröffentlichung,

[60] http://repositories.webometrics.info/, gelesen 18.1.2011.

- das Angebot von Zusatzdiensten und
- klare Vereinbarungen zu Rechten und Pflichten aller Beteiligter,

helfen, den Bestand an Publikationen und damit den Inhalt zu vergrößern.

Der dritte Aspekt ist die technische Qualität des IR, für deren Einhaltung und Verbesserung ausschließlich die Betreiber verantwortlich sind. Die technische Qualität zeichnet sich durch

- Wahrung der Integrität und Langzeitarchivierung aller Komponenten der Publikationen,
- Herstellung größtmöglicher Verfügbarkeit und Sicherheit des Dienstes und
- Schaffung der Voraussetzungen für die Sichtbarkeit im Internet

aus.

Qualitätskriterien, die sowohl den inhaltlichen als auch den technischen Aspekt berücksichtigen, sind im DINI-Zertifikat[61] zusammengefasst. Durch die Formulierung von Kriterien wird ein Standard geschaffen, an dem sich Betreiber bei Konzeption, Aufbau und Entwicklung eines OA-Repositories orientieren können. OA-Repositories erhalten das DINI-Zertifikat, wenn die darin enthaltenen Mindestanforderungen erfüllt und damit die Voraussetzungen vorhanden sind, in eine gemeinsame Infrastruktur aller deutschen und internationalen OA-Repositories einbezogen zu werden. Neben den Mindestanforderungen werden Empfehlungen gegeben, deren Befolgung auf künftige Entwicklungen vorbereiten sollen. Die Schaffung der Voraussetzungen für eine gute Sichtbarkeit des Gesamtangebotes und der Publikationen im Internet nimmt eine zentrale Rolle unter den Kriterien ein. Das erste der Kriterien des DINI-Zertifikats 2010 (DINI 2010a), „Sichtbarkeit des Gesamtangebotes“, beinhaltet die Bekanntmachung des Angebotes in der betreibenden Einrichtung und die Anmeldung des Dienstes bei DINI und OpenDOAR, womit die Aufnahme in entsprechende Web-Verzeichnisse verbunden ist. Die Registrierung bei Nachweisdiensten wie DRIVER und als Datenprovider der Open Access Initiative wird empfohlen. Das Kriterium „Erschließung und Schnittstellen“ enthält Festlegungen für die inhaltliche und formale Erschließung der Publikationen und macht eine OAI-Schnittstelle zur Pflicht. OA-Repositories, die diesen Kriterien genügen, bieten damit die Voraussetzungen für die Sichtbarkeit und damit für die Nutzung der Publikationen.

Die Einhaltung des Qualitätsstandards sagt allerdings noch nichts darüber aus, wie gut die Sichtbarkeit ist. Eine der Hauptaufgaben eines IR besteht darin, Wissen zu verbreiten, was nur durch gute Sichtbarkeit erreicht wird. Auskunft über die Sichtbarkeit des gesamten Repository kann der Web Impact Factor (WIF) oder das “Ranking Web of World Repositories” geben, in welches der WIF eingeht. Hier steht der Vergleich mit anderen Repositories im Vordergrund. Hinweise auf die Sichtbarkeit der Publikationen in viel detaillierterer

[61] Bisher existieren DINI-Zertifikate der Jahre 2004, 2007 und 2010, in denen die Standards an die aktuelle Entwicklung angepasst wurden.

Form können Betreiber durch die Analyse der Nutzungsdaten[62] mit NoRA erhalten. So kann ermittelt werden, wo an der Verbesserung der Sichtbarkeit gearbeitet werden muss, um eine höhere Nutzungshäufigkeit zu erzielen und damit die Qualität des OA-Repositoriy insgesamt zu erhöhen.

[62] Die Erhebung und Veröffentlichungvon Nutzungsdaten von Publikationen ist ein Zusatzdienst, der im DINI-Zertifikat 2010 empfohlen wird.

2 Downloadzahlen von Open Access Repositories

2.1 Ermittlung von Downloadzahlen aus Webserver-Logfiles

2.1.1 Was ist ein Logfile?

Beim Betrieb eines Webservers ensteht bei jedem Zugriff auf eine Website ein Datensatz in Form von ASCII[63]-Zeichen, in welchem mehr oder weniger detaillierte Angaben zum Zugriff gespeichert werden. Diese Datensätze bilden Logfiles.

Der Inhalt, welcher bei fast allen Servern aus NCSA[64]-Standardelementen besteht, ist von der Serversoftware abhängig und kann ausserdem konfiguriert werden. Das am häufigsten verwendete Format ist das Combined Log Format, welches von der Grundkonfiguration des Apache-Webservers[65] erzeugt wird. Vom Apache-Webserver können verschiedene Arten von Logfiles wie Access Log, Error Log, Script Log usw., die unterschiedliche Ereignisse protokollieren, generiert werden. Da im Zusammenhang mit Downloadzahlen die Zugriffe auf den Webserver interessieren, ist im Folgenden mit Logfile immer das File Access Log gemeint. Hier das Beispiel eines Datensatzes des Logfiles:

```
141.20.8.170 - - [09/Nov/2006:14:47:43 +0100] "GET /e_talks/downloads/cms_051213/metadaten.pdf HTTP/1.1" 206 131072 "http://www.google.de/search?q=Sabine+Henneberger+edoc&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:de:official&client=firefox-a" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"
```

Die Struktur wird in der folgenden Tabelle erläutert[66]. Als Client wird die Kombination aus Host und Agent bezeichnet.

Tab. 1: Struktur eines Datensatzes des Logfiles

Nr.	Inhalt	Bedeutung	Bezeichnung
1	141.20.8.170	IP-Adresse des Clients (numerisch oder als DNS-Name)	Host[67]
2	-	Identifikation des Client-Computers (in der Regel nicht vorhanden)	
3	-	Identifikation des Nutzers (bei Authentifizierung)	

[63] ASCII: American Standard of Information Interchange.

[64] NCSA: National Center for Supercomputing Applications, http://www.ncsa.uiuc.edu/, gelesen 20.1.2011.

[65] Siehe http://httpd.apache.org/docs/logs.html, gelesen 20.1.2011.

[66] Genaue Angaben und Erläuterungen zum Inhalt von Logfiles findet man z. B. in (Heindl 2003).

[67] In der Literatur wird auch oft die Bezeichnung „Remote Host“ verwendet.

4	[09/Nov/2006:14:47:43 +0100]	Zeit der Anfrage (mit Abweichung von GMT[68])	Time
5	"GET /e_talks/downloads/cms_051213/ metadaten.pdf HTTP/1.1"	Anfrage des Clients (Zugriffsmethode, Pfad zum File, Name und Version des Protokolls)	Request
6	206	Erfolgsstatus der Anfrage	Status
7	131072	Anzahl der übertragenen Bytes bei erfolgreicher Anfrage	
8	"http://www.google.de/search?q=Sabine+ Henneberger+edoc&ie=utf-8&oe=utf-8&aq =t&rls=org.mozilla: de:official&client=firefox-a"	URL, von der aus der Client weitergeleitet wurde	Referrer
9	"Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"SV1; .NET CLR 1.1.4322)"	Beschreibung des Clients (Browser, Betriebssystem)	Agent[69]

Andere Formate sind:

- Common Logfile Format: Besitzt nur die Positionen 1-7
- W3C[70] Extended Format: enthält eine Reihe zusätzlicher Positionen[71]
- Microsoft IIS Format: Positionen 1-7, Trennung duch Kommas[72,73]

2.1.2 Der formale Inhalt eines Logfiles

Jede von einem Client ausgelöste Aktion wird in einem Datensatz - einem Hit[74] - protokolliert, egal ob erfolgreich oder nicht. So enthält das Logfile, welches alle Anfragen an den edoc-Server der Humboldt-Universität für einen Tag speichert, beispielsweise für den 30.6.2006 insgesamt 133 385 Datensätze. Diese Zahl kann nicht mit der Anzahl der Aktionen von Nutzern, also Personen, die am Inhalt von Dokumenten interessiert sind, gleichgesetzt werden, da einerseits pro Zugriff durch einen Nutzer mehrere Datensätze generiert werden können, andererseits aber alle durch Robots ausgelöste Hits, auf die später genauer eingegangen

[68] GMT: Greenwich Mean Time.

[69] In der Literatur wird auch oft die Bezeichnung „User Agent“ verwendet.

[70] W3C: World Wide Web Consortium, http://www.w3.org/, gelesen 20.1.2011.

[71] http://www.w3.org/TR/WD-logfile.html, gelesen 20.1.2011.

[72] IIS ist ein Akronym für Microsofts Internet-Informationsserver.

[73] Bei den Akronymen für Formate gibt es in der Literatur einige Verwirrung. So wird mit CLF sowohl das Common Logfile Format als auch das Combined Logfile Format bezeichnet, welches die Positionen 1-9 wie in der Tabelle enthält. An anderer Stelle wird dieses Format ECLF (Extended Common Logfile Format) genannt. Bei den weiteren Beispielen handelt es sich um das Format ECLF mit den Positionen 1-9.

[74] Die Begriffe Hit und Request werden von Analyse-Programmen synonym benutzt. Korrekt ist die Verwendung von Hit zur Bezeichnung eines Datensatzes im Logfile. Der Request ist ein Bestandteil des Hits.

wird, in der gleichen Weise protokolliert werden. Beim Aufruf einer Website, die mehrere Bilder enthält, wird z. B. ein Datensatz für das HTML-File selbst und für jedes Bild ein weiterer Datensatz geschrieben. Wird die Website über ein css-Stylesheet[75] formatiert, entsteht für dieses auch ein Datensatz. So können durch den Aufruf einer einzigen Website eine Vielzahl von Datensätzen erzeugt werden. Ein Zugriff auf ein PDF-File erzeugt eine Anzahl von Datensätzen, die von der Größe des Dokumentes abhängig ist. In der Regel wird bei Zugriff auf ein PDF-File das Dokument nur unvollständig, also mit Status Code 206, in mehreren Abschnitten übertragen. Wurde das File vollständig übertragen, entsteht ein Datensatz mit Status Code 200. Protokolliert werden ebenfalls erfolglose Zugriffe, etwa der Aufruf einer nicht existierenden Website. Nur ein Teil der Datensätze protokolliert eine Anfrage im Sinne eines Downloads und die Anzahl der Datensätze pro Download ist abhängig davon, um welchen Typ von Website es sich handelt.

Hier ein konkretes Beispiel für eine Sequenz von Datensätzen des Logfiles:

141.20.8.170 - - [09/Nov/2006:14:47:43 +0100] "GET /e_talks/downloads/cms_051213/metadaten.pdf HTTP/1.1" 206 131072 "http://www.google.de/search?q=Sabine+Henneberger+edoc&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:de:official&client=firefox-a" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

141.20.8.170 - - [09/Nov/2006:14:47:43 +0100] "GET /e_talks/downloads/cms_051213/metadaten.pdf HTTP/1.1" 206 812022 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

Nachdem in Google “Sabine Henneberger” als Suchbegriff eingegeben wurde, erfolgt der Zugriff auf metadaten.pdf durch einen Nutzer, der die IP-Adresse 141.20.8.170 und den Browser Mozilla Firefox benutzt. Das PDF-Dokument wird in Abschnitten heruntergeladen, aber beide Male nicht vollständig, was am Status-Code 206 erkennbar ist.

141.20.8.170 - - [09/Nov/2006:14:49:52 +0100] "GET /e_info/partner.php HTTP/1.1" 200 22779 "http://www.google.de/search?q=Sabine+Henneberger+Ansprechpartner&btnG=Suche&hl=de&client=firefox-a&rls=org.mozilla%3Ade%3Aofficial&hs=Zoo" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

141.20.8.170 - - [09/Nov/2006:14:49:52 +0100] "GET /e_images/spacer.gif HTTP/1.1" 200 43 "http://edoc.hu-berlin.de/e_info/partner.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

141.20.8.170 - - [09/Nov/2006:14:49:52 +0100] "GET /e_images/hub_logo2.jpg HTTP/1.1" 200 31171 "http://edoc.hu-berlin.de/e_info/partner.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

[75] Gemeint ist damit ein File, welches Angaben zur Formatierung einer Website enthält.

141.20.8.170 - - [09/Nov/2006:14:49:52 +0100] "GET /e_images/dot.gif HTTP/1.1" 200 43 "http://edoc.hu-berlin.de/e_info/partner.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

Von der gleichen IP-Adresse aus wurde nach "Sabine Henneberger Ansprechpartner" gesucht und anschließend auf partner.php zugegriffen. Die Seite wurde vollständig geladen (Status-Code 200). Da die Website mehrere Bilder enthält, wird für jedes einzelne davon auch ein Datensatz erzeugt. Der Referrer ist jeweils partner.php.

141.20.8.170 - - [09/Nov/2006:14:50:28 +0100] "GET /e_autoren/index.php HTTP/1.1" 200 18173 "http://edoc.hu-berlin.de/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

141.20.8.170 - - [09/Nov/2006:14:50:32 +0100] "GET /e_autoren/alternativen.php?arbeit=Dissertationen%20%C2%BB&index=index.php&nav=diss HTTP/1.1" 200 17101 "http://edoc.hu-berlin.de/e_autoren/index.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.8) Gecko/20071008 Firefox/2.0.0.8"

Auf der Website http://edoc.hu-berlin.de wird der Link zu e_autoren/index.php geklickt und man gelangt von dort aus zu alternativen.php

2.1.3 Analyse von Logfiles

Man sieht bereits an diesem sehr kurzen Beispiel, wieviel verschiedene Informationen in den Logfiles enthalten sind, die Antwort auf die unterschiedlichsten Fragestellungen zum Verhalten von Nutzern geben können. Solche Fragestellungen sind Gegenstand des Web Usage Mining, einem Teilbereich des Web Mining, in welchem Verfahren des Data Mining angewendet werden, um Strukturen im Internet aufzudecken[76]. Das Zählen der Zugriffe von Nutzern und damit auch der Downloads von Publikationen gehört dazu. Als Downloadzahlen sollen die Zugriffe durch menschliche Nutzer, im Folgenden der Kürze wegen als Human User bezeichnet, erfasst werden.

2.1.3.1 Standardisierung der Logfileanalyse

Die Hauptakteure unter den Anbietern von Online-Publikationen wie Herausgeber und Bibliothekare haben verständlicherweise großes Interesse an Informationen über die Nutzung ihrer Produkte und Bibliotheksbestände, die aus den Logfiles gewonnen werden können. Um solche Informationen verlässlich und austauschbar zu machen, muss ihre Gewinnung einheitlichen Voraussetzungen genügen und die Darstellung vergleichbar sein. Zu diesem Zweck wurde COUNTER (Counting Online Usage of Networked Electronic Resources)[77] als Non-Profit-Organisation gegründet. Im COUNTER Code of Practice wird festgelegt, welche

[76] Eine Übersicht dazu findet man z.B. bei (Heindl 2003).

[77] http://www.projectcounter.org/, gelesen 20.1.2011.

Angaben Nutzungsstatistiken, die sogenannten Usage Reports, enthalten sollen und wie diese Angaben gewonnen werden. Genaue Bestimmungen gibt es auch für das Format der Usage Reports. Zusammengefasst werden diese Richtlinien (aktueller Stand Januar 2011) im

- COUNTER Code of Practice for Journals and Databases: Release 3 (COUNTER 2008) und
- COUNTER Code of Practice for Books and Reference Works: Release 1 (COUNTER 2006).

Eingeflossen sind die bereits vorher verfassten ICOLC Guidelines for Statistical Measures of Usage of Web-based Information Resources[78]. Die Einhaltung der Richtlinien muss jährlich überprüft werden, um den Status „COUNTER-compliant", also „COUNTER-konform", für die Nutzungsstatistik eines Anbieters aufrecht zu erhalten[79].

Damit können Nutzungsstatistken verschiedener Anbieter, die COUNTER-konform sind, verglichen werden. Im Code of Practice ist z. B. festgelegt, dass nur die Hits mit Status Code 200 und 304 relevant sind. Doppelklicks sind als Hits vom gleichen Nutzer mit zeitlichem Abstand unter 10 Sekunden definiert und werden als 1 Hit gezählt. Die Festlegungen von COUNTER gelten inzwischen als Pseudostandard bei der Erzeugung von Nutzungsstatistiken. Über das SUSHI-Protokoll[80] ist eine automatische Erstellung der COUNTER-konformen Nutzungsstatistik möglich.

Wurde anfangs davon ausgegangen, dass die COUNTER-konformen Nutzungsstatistiken in abgeschlossenen Systemen erhoben wurden, bei denen jeder Nutzer sich vorher identifizieren muss, und automatisch erzeugte Zugriffe nicht vorkommen, wurden in der letzten Ausgabe des Code of Practice for Journals and Databases 2008 Bedingungen für den Ausschluss von Robots aufgenommen. Offensichtlich kommen aber nur wenige Robots in Betracht, da die zur Identifizierung von Robots bei COUNTER veröffentlichten Listen vergleichsweise kurz sind.

Bei OA Repositories ist das anders, da sie für jeden Zugriff offen sind. Ein erheblicher Anteil an Hits, oft 50% und mehr, wird automatisch durch Computerprogramme, die das Internet nach Informationen durchsuchen, den Robots[81], erzeugt. Das sind unter anderem Suchmaschinen, die Websites suchen und in ihre Indices aufnehmen, aber auch z. B. Programme, die das Internet nach bestimmten Inhalten von Websites wie E-Mail-Adressen durchsuchen. Diese Robot Hits müssen erkannt und ausgeschlossen werden. Ein Teil der Robots

[78] ICOLC: International Coalition of Library Consortia, http://www.library.yale.edu/consortia/2001webstats.htm, gelesen 24.1.2011.

[79] Andere Standards sind die von LogEC und IFABC (International Federation of Audit Bureaux of Circulations), http://logec.repec.org/ und http://www.ifabc.org/ , gelesen 24.1.2011.

[80] SUSHI: Standardised Usage Statistics Harvesting Initiative.

[81] Spider, Crawler und Spammer sind einerseits Bezeichnungen für spezielle Robots. Andererseits werden diese Begriffe aber einfach als Synonym für Robot, oder kurz Bot, verwendet.

gibt sich in den Logfiles zu erkennen, ein anderer Teil, der oft bewusst nicht erkannt werden will, kann nur anhand von Mustern im Logfile entdeckt werden.

2.1.3.2 Bildung von Sessions

Bevor der Ausschluss von Robots behandelt werden kann, müssen zunächst die Begriffe „Session“ und „Sessionizing“ erläutert werden, soweit es für die Identifikation von Robot Hits notwendig ist. „Eine Sitzung (Session) ist ein zeitlich in engem Zusammenhang stehender Abruf von Dokumenten durch einen Besucher.“, so die Definition von Heindl (Heindl 2003).

Durch die Verwendung von Proxy-Servern[82] und die dynamische Vergabe von IP-Adressen von Internet-Providern ist es nicht generell möglich, einen Nutzer anhand der IP-Adresse von einem anderen zu unterscheiden, d. h. unter einer IP-Adresse können sich mehrere Nutzer verbergen. Andererseits können einem Nutzer durch einen Provider mehrere IP-Nummern zugeteilt gewesen sein. Eine Möglichkeit, einen Nutzer korrekt zu identifizieren, wäre die Identifikation der benutzten Browsers-Installation durch das Setzen von Cookies[83], was aber nur zum Teil praktiziert wird. Eine andere Variante ist die Bildung von Sessions, genannt Sessionizing, anhand der Logfiles, die allerdings nicht die gleiche Zuverlässigkeit besitzt, da sie von Grundannahmen ausgehen muss, die nicht unbedingt zutreffen.

Eine Session wird anhand der Kombination von Host und Agent definiert. Eine neue Session beginnt, wenn eine Kombination erstmalig auftritt oder wenn eine bestimmte Zeit seit ihrem letzten Auftreten vergangen ist. Ein Pseudostandard für diesen Zeitraum sind 1800 Sekunden[84]. D. h. bei allen Datensätzen des Logfiles, die die gleiche Kombination von IP-Adresse und Browser besitzen und für deren Zeit gilt Start Time <= Time <= End Time[85], geht man davon aus, dass sie zum gleichen Nutzer gehören und wegen des zeitlichen Zusammenhanges eine Session bilden. Sessionizing wird aus folgenden Gründen betrieben:

- Bestimmung unterschiedlicher Nutzer
- Ermittlung von Nutzerverhalten
- Ermittlung von Robots anhand von Verlaufsmustern

Für diese Arbeit ist nur der letzte Punkt von Bedeutung, da die Summierung der Zugriffe durch Human Users das Ziel ist. Die Bestimmung von unterschiedlichen Nutzern und das Nutzerverhalten sind hier nicht relevant, da solche Informationen in den Nutzungsdaten von IR nicht enthalten sind.

---

[82] Proxy-Server: Zwischenspeicher für bereits aufgerufene Websites.

[83] Browser-Cookie: Informationen, die vom Server an den Client gesendet werden und bei weiteren Aufrufen des Servers an diesen wieder übertragen werden (siehe auch http://de.wikipedia.org/wiki/HTTP-Cookie, gelesen 20.1.2011).

[84] Dieser Wert entspricht dem COUNTER-Standard.

2.1.3.3 Ausschluss von Robots

Woran kann oder könnte man einen Robot Hit erkennen?

1. Request auf robots.txt
   Das File enthält eine Liste von Websites, die nicht in den Index von Suchmaschinen aufgenommen werden sollen. Gemäß dem Robot Exclusion Standard[86] sollten Robots immer dieses File lesen und berücksichtigen. Dieser Standard wird aber nicht von allen Robots befolgt.
2. Agent
   Entsprechend der Direktive von Eichmann (Eichmann 1995) sollen sich Robots durch ihre Browserbezeichnung offenbaren und diese im Agent mitteilen, was von vielen Robots, aber wiederum nicht von allen befolgt wird. Es gibt Databases im Internet, die diese Robot Agents enthalten, aber nicht vollständig und aktuell sein können.
3. Request-Methode
   Ein Human User wird in der Regel die GET-Methode[87] verwenden, es sei denn, ein Proxy-Server ist dazwischen geschaltet und verwendet die HEAD-Methode[88]. Die ausschließliche Verwendung von HEAD weist auf einen Robot hin.
4. Referrer
   Der Referrer bleibt leer, wenn ein Hit durch eine Bookmark oder direkte Eingabe der URL durch einen Human User entsteht. Ist bei allen Hits einer Session oder eines Clients der Referrer leer, ist das ein Hinweis auf einen Robot.
5. DNS-Name oder IP-Adresse des Host
   Wandelt man IP-Adressen in DNS-Namen um, erhält man oft schon aus diesem Namen den Hinweis auf einen Robot. Über WhoIs-Server[89] kann man weitere Informationen erhalten. Wie für Agents existieren Listen im Internet, die IP-Adressen oder DNS-Namen bekannter Robots enthalten. Diese Databases vollständig und aktuell zu halten ist, wie für Robot Agents, unmöglich.

---

[85] Start Time bezeichnet den Zeitpunkt des ersten, End Time den Zeitpunkt des letzten Auftretens der Kombination Host-Agent.

[86] Der Robot Exclusion Standard ist kein Standard im eigentlichen Sinn, sondern eine Übereinkunft einer Gruppe von Suchmaschinenentwicklern aus dem Jahr 1994.

[87] GET-Methode: Anforderung einer HTML-Seite (Website).

[88] HEAD-Methode: Anforderung des HEAD-Teils einer HTML-Seite.

[89] WhoIs-Server enthalten freiwillige Angaben von Serverbetreibern zur Institution.

6. Muster der Session, in der sich der Hit befindet
   Viele Robots erzeugen Muster im Logfile, z. B. durch gleiche zeitliche Abstände, Häufigkeit und Regelmäßigkeit des Auftretens im Tagesverlauf. Der Zugriff nur auf PDF-Files oder nur auf Bilder oder nur auf HTML-Files, ohne auf enthaltene Bilder zuzugreifen, ist ebenfalls ein Hinweis auf einen Robot.

Ein Ausschluss anhand der Merkmale 1-5 ist durch relativ einfache Algorithmen realisierbar. Der Erfolg hängt weitgehend davon ab, wie gut die Informationen über existierende Robots sind. Die Bildung von Sessions und die Identifikation als Robot-Session aufgrund bestimmter Muster ist wesentlich komplizierter. Ein Tool zum Sessionizing und darauf basierender Identifikation von Robots, das Robot Detection Tool, und die Anwendung als Preprocessing für die Logfileanalyse stellen Bomhardt et al. vor (Bomhardt et al. 2005). Geens et al. vergleichen in einer Studie verschiedene Techniken zur Robot-Erkennung und schlagen eine neue kombinierte Methode vor, die angeblich bis zu 90% der Robots entdecken kann (Geens et al. 2006). Einen völlig anderen, wahrscheinlichkeitstheoretischen Ansatz zur Robot-Erkennung verfolgen Stassopouloua und Dikaiakos (Stassopouloua und Dikaiakos 2009). In den verbreiteten Analyseprogrammen werden diese Methoden jedoch noch nicht verwendet.

Zusammenfassend gilt, dass ein großer Teil von Robot Hits als solche zweifelsfrei identifiziert werden können. Bei den übrigen kann man aufgrund der vagen Merkmale nicht mit Sicherheit entscheiden, ob es sich um einen Robot oder Human User handelt. Man kann nur versuchen, eine Methode zu verwenden, die den Ausschluss von Robots optimiert. Erschwert wird die Erkennung von Robot Hits durch bewusste Tarnung und ständig neu auftretende Robots. Listen von Robot Agents, IP-Nummern oder DNS-Namen von bekannten Robots existieren zwar, sind aber nie aktuell. Die Robots des deutschsprachigen Raums sind in solchen Listen unterrepräsentiert.

2.1.4 Tools zur Logfileanalyse

Es gibt eine ganze Reihe an Tools zur Analyse von Logfiles, teils Freeware oder Shareware und Open Source[90], und natürlich kostenpflichtige Tools zur Installation auf dem eigenen Server. Man kann auch Dienstleistungen in der Form von externer Auswertung über das Internet in Anspruch nehmen.

Hier soll es um die Analyse der Logfiles mit Hilfe von Tools gehen, die auf dem Server installiert werden, insbesondere frei verfügbare, da vor allem diese für OA Repositories genutzt werden. Um eine passende Auswahl für den edoc-Server zu treffen, wurde im Jahr 2007 ein Vergleich von in Frage kommender Tools durchgeführt. Dazu wurden die zwei bekanntesten Tools AWStats und Analog mit großen Communities, das weniger bekannte W3Perl mit einer kleineren Community und das als sehr flexibel geltende, meist in akademischen Kreisen verwendete WUMprep ausgewählt. WUM ist ein Tool, das über die Logfileanalyse durch Filterung, Roboterkennung und Sessionizing hinausgeht und Klassifizierungen von Webinhalten ermöglicht

[90] Die Begriffe „Freeware“, „Shareware“ und „Open Source“ werden z. B. hier definiert und voneinander abgegrenzt: http://opensourcestrategies.blogspot.com/2005/09/freeware-vs-shareware-vs-open-source.html, gelesen 21.1.2011.

(Dettmar 2004). Hier ist aber nur die Vorbereitung der Logfiles mit WUMprep zur Zählung von Zugriffen von Interesse. Um zu überprüfen, ob sich auch die Lightversion eines kommerziellen Tools eignet, wurde Deep Web Analyzer Light einbezogen. Die folgenden Versionen wurden überprüft:

- AWStats[91], Version 6.7
- W3Perl[92], Version 3.0
- Analog[93], Version 6.0
- WUMprep[94], Version 0.10.0
- Deep Web Analyzer Light[95], Version 3.2

Der Vergleich brachte wichtige Erkenntnisse über die Zuverlässigkeit von Logfileanalysen und ist deshalb Bestandteil dieser Arbeit geworden.

2.1.4.1 Logfiles und Datenschutz

OA Repositories gelten als Content Provider, da sie im Gegensatz zu Access Providern Inhalte anbieten und nicht nur den technischen Zugang zum Internet ermöglichen. Deshalb wird auf OA Repositories das Telemediengesetzes (TMG)[96], welches seit Februar 2007 gilt, angewendet. Darin wird unter anderem der Umgang mit Nutzungsdaten geregelt. Laut einem Urteil des Landgerichts Berlin[97] ist die Speicherung von IP-Adressen unzulässig. Die Auffassung, dass IP-Adressen personenbezogene Daten sind, wurde in juristischen Kreisen zwar nicht uneingeschränkt geteilt, das Gesetz besteht aber seither unverändert (Kaufmann 2007). Damit entfallen viele Möglichkeiten, Logfiles nach unbekannten Robots zu durchsuchen und sie zu identifizieren. Die vergleichende Untersuchung von Tools, so wie sie im Folgenden vorgestellt wird, wäre heute nicht mehr gestattet, da die Ergebnisse anhand der vollständigen Logfiles überprüft wurden. Dieser Teil der Arbeit wurde schon 2007 mit einem Logfile aus dem Jahr 2006 durchgeführt. Die Ergebnisse können wegen des TMG leider nicht für die aktuellen Programmversionen und aktuelle Logfiles verifiziert werden. Man kann aber davon ausgehen, dass die Algorithmen zur Robot-Erkennung nicht wesentlich geändert wurden.

[91] http://awstats.sourceforge.net/, gelesen 21.1.2011.

[92] http://www.w3perl.com/, gelesen 21.1.2011.

[93] http://www.analog.cx/, gelesen 21.1.2011.

[94] http://sourceforge.net/projects/hypknowsys/files/WUMprep/, gelesen 21.1.2011.

[95] http://www.deep-software.com/, gelesen 21.1.2011.

[96] Telemediengesetz vom 26. Februar 2007 (BGBl. I S. 179).

[97] Urteil vom 6.9.2007, AZ. 23 S 3/07.

Da sich die Anzahl der Robots seither stark vergrößert hat, ist zu erwarten, dass die Ergebnisse heute noch heterogener ausfallen als damals[98].

2.1.4.2 Vorgehensweise beim Vergleich

Als Datenbasis wurde das Logfile eines Tages gewählt. Dieses Testfile umfasst 133 385 Zeilen (Hits) und zeichnet die Zugriffe vom 30.6.2006, 01:51 bis 01.07. 2006, 01:38 auf. Eine Zeile ist wegen fehlendem Status Code ungültig.

Dieses Logfile wurde mit den fünf Tools analysiert, wobei die Konfigurationen der Tools so weit wie möglich gleich waren. Da sich die Konfigurationsmöglichkeiten unterscheiden, konnten nicht völlig gleiche Ausgangsbedingungen hergestellt werden. In allen Auswertungen wurden nur Hits mit Status Code 200 und 304[99] einbezogen. Zur Ermittlung der Robots wurde das zum Tool gelieferte bzw. empfohlene File benutzt, welches eine Liste von bekannten Robots enthält. Die Algorithmen der Tools zur Robot-Erkennung unterscheiden sich. Zur Kontrolle der Datenbasis wurde das Programm SPSS[100] verwendet. Es wurden nur die Eigenschaften der Tools verglichen, die für die Ermittlung von Downloadzahlen wichtig sind. Folgende Kriterien wurden überprüft:

1. Konfigurierbarkeit
   Kann das Programm auf die gewünschte Art der Analyse und die speziellen Gegebenheiten des Servers angepasst werden?
2. Erkennung und Ausschluss von Robots
   Wie zuverlässig ist die Erkennung von Robots?
3. Zählweise der Hits und Auswertung des Status
   Welche Hits werden summiert und kann die Art der Summierung verändert werden?
4. Plausibilität der Angaben
   Sind die Angaben im Ergebnis nachvollziehbar oder gibt es Unklarheiten, welche Datensätze wie einbezogen werden?
5. Sessionizing
   Wird es angeboten, hat es Auswirkung auf die Zugriffszahlen und ist es brauchbar?
6. Ausgabeform der Ergebnisse
   Hier interessiert nicht die vom Programm generierte Website, sondern ob man die Ausgabe so gestalten kann, dass eine spätere Zusammenführung mit der Metadatenbank möglich wird.

---

[98] Aktuelle Versionen der Programme am 21.1.2011: AWStats 6.0 vom 5.12.2010, W3Perl 3.11 vom 8.1.2011, Analog 6.0 vom 19.12.05, WUMprep 0.10.0 vom 21.8.2005, Deep Web Analyzer heißt jetzt Deep Log Analyzer 4.0.

[99] 200: Das File wurde erfolgreich vollständig heruntergeladen; 304: Der Client kann auf eine unveränderte Kopie in seinem Cache zurückgreifen.

7. Ergebnisse

   Stimmen die Ergebnisse mit denen aus der SPSS-Analyse überein?

Das Ergebnis für eine URL ist die Anzahl der Zeilen des Logfiles, die einen Zugriff auf diese URL repräsentieren und nicht durch Algorithmus oder Konfiguration ausgeschlossen wurden. Wird durch eine URL ein PDF-File markiert, ist das Ergebnis die Downloadzahl dafür. Ausgeschlossen werden Zeilen, die nicht den Status Code 200 oder 304 besitzen. Die folgende Tabelle zeigt die Häufigkeiten der Status Codes im Testfile.

Tab. 2: Häufigkeiten von Status Codes im Testfile

Status Code		**Häufigkeit**
Gültig	200	95867
	206	17917
	301	2671
	302	3091
	304	12274
	400	34
	403	18
	404	1509
	408	1
	500	2
	Gesamt	133384
Fehlend	System	1
Gesamt		133385

Nur Zeilen mit 200 und 304 werden ausgewertet, die übrigen sind für die Ermittlung der Zugriffszahlen nach COUNTER-Standard uninteressant. Analysiert werden demnach 108.157 Zeilen des Testfiles. Um die Ergebnisse für eine Stichprobe zu überprüfen, wurden drei Publikationen, auf die an diesem Tag zugegriffen wurde, willkürlich ausgewählt. Hierbei geht es nicht um eine repräsentative Stichprobe, sondern nur um die Demonstration der unterschiedlichen Ergebnisse. Für die Volltexte seifarth.pdf, Graefe.pdf und Zimmer.pdf wurden die Zugriffszahlen mit dem Programm SPSS direkt aus dem Logfile ermittelt, aus dem nur die Zeilen selektiert wurden, die diese Filenamen enthielten und den Status Code 200 oder 304 aufwiesen, was überschaubare 43 Zeilen ergab.

[100] Seinerzeit wurde SPSS in der Version 14 verwendet.

Tab. 3: Requests auf die Beispielfiles

Request	Häufigkeit
GET /dissertationen/seifarth-joerg-2004-07-07/PDF/seifarth.pdf HTTP/1.0	27
GET /dissertationen/seifarth-joerg-2004-07-07/PDF/seifarth.pdf HTTP/1.1	8
GET /habilitationen/graefe-michael-2001-07-17/PDF/Graefe.pdf HTTP/1.0	4
GET /habilitationen/graefe-michael-2001-07-17/PDF/Graefe.pdf HTTP/1.1	1
GET /habilitationen/zimmer-claus-2001-04-10/PDF/Zimmer.pdf HTTP/1.0	2
HEAD /habilitationen/graefe-michael-2001-07-17/PDF/Graefe.pdf HTTP/1.1	1 (a)
Gesamt	43

Tab. 4: Agents der Requests auf die Beispielfiles

Graefe.pdf	Häufigkeit
-	1 (b)
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)	1
Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	1 (c)
Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)	1 (d)
Mozilla/5.0 (Windows; U; Windows NT 5.1; de-DE; rv:1.7.5) Gecko/20041108 Firefox/1.0	2
Gesamt	6
seifahrt.pdf	**Häufigkeit**
MnogoSearch/3.2 (+http://www.cms.hu-berlin.de/portale/entwickler/mnogosearch/index.html)	1 (e)
Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)	1
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)	1
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Crazy Browser 1.0.5; .NET CLR 1.1.4322)	1
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)	25
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; Arcor 5.005)	1
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; freenet.de FRNKDZ 5.0)	1
Mozilla/5.0 (Windows; U; Windows NT 5.1; de-DE; rv:0.9.4) Gecko/20011019 Netscape6/6.2	1
Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.0.4) Gecko/20060508 Firefox/1.5.0.4	2
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7.5) Gecko/20041107 Firefox/1.0	1
Gesamt	35

Zimmer.pdf	Häufigkeit
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322)	2

Ausgeschlossen wurden:

- für Graefe.pdf 1 Zugriffe wegen (a) Methode HEAD und dem leeren Agent (b) und 2 Zugriffe wegen Robot Agents (c) und (d)
- für seifahrt.pdf 1 Zugriff wegen Robot-Agent (d)

Damit ergeben sich als Downloads aus der SPSS-Analyse:

- Graefe.pdf: $D_G = 3$
- seifahrt.pdf: $D_S = 34$
- Zimmer.pdf: $D_Z = 2$

In die Auswertung sollen nur die Zeilen mit Status Code 200 und 304 einbezogen werden. In den folgenden Abschnitten sind die Testergebnisse für die fünf Tools zusammengefasst.

2.1.4.3 AWStats

Tab. 5: Bewertung von AWStats

Konfigurierbarkeit	Sehr gut, Status Codes können beliebig ein- oder ausgeschlossen werden, die Angabe von Hosts und Requests, die ausgeschlossen werden sollen, ist möglich.
Robots	Anzahl 47, Erkennung über Listen[101] durch Vergleich mit Agent
Zählweise	Es werden nur Zeilen mit Status 200 und 304, wie konfiguriert, gezählt.
Plausibilität	Nein, siehe Abbildung 3 und erklärenden Text
Sessionizing	Vorhanden, wird aber nicht zur Robot-Erkennung, sondern zur Bildung von Visits genutzt.
Ausgabe Ergebnisse	Es gibt gut geeinete Ausgaben in den Formaten txt, html und xhtml.
Ergebnisse	• Graefe.pdf: $D_G = 0$ • seifahrt.pdf: $D_S = 34$ • Zimmer.pdf: $D_Z = 0$

[101] Eine Liste robots.pm mit Robot Agents ist im Programmpaket enthalten.

Die Angaben im Summary, welches die folgende Abbildung zeigt, oder an einer anderen Stelle sind nicht korrekt. Das kann überprüft werden, wenn man die Angaben im Summary mit den tatsächlichen Werten vergleicht.

Summary

Reported period	Year 2006			
First visit	30 Jun 2006 - 01:51			
Last visit	01 Jul 2006 - 01:38			
	Unique visitors	Number of visits	Pages	Hits
Viewed traffic *	<= **5646** Exact value not available in 'Year' view	**6268** (1.11 visits/visitor)	**15218** (2.42 Pages/Visit)	**65701** (10.48 Hits/Visit)
Not viewed traffic *			**64116**	**66194**

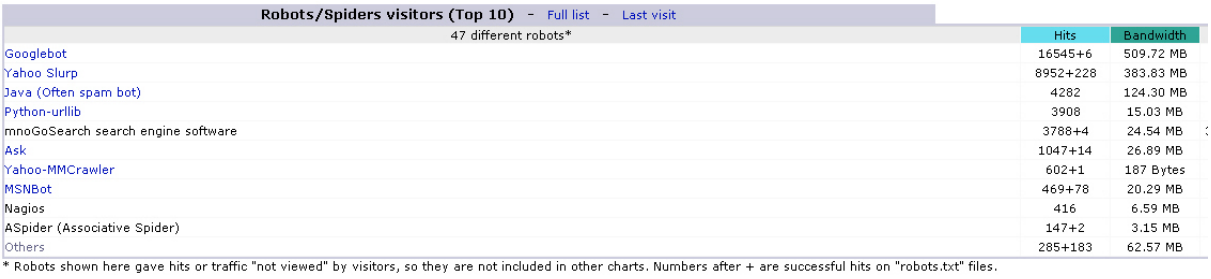
* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Abb. 1: AWStats Summary

Viewed traffic (vt) ist die Summe aller Hits, also Zeilen des Testfiles, die bei der Berechnung der Downloadzahlen berücksichtigt werden, not viewed traffic (nvt) ist die Summe aller Hits, die ausgeschlossen werden. Addiert man die Angaben aus dem Summary, ergeben sich 131.895 Hits und damit weniger Zeilen, als im Testfile vorhanden sind. Berechnet man vt und nvt aus den einzelnen Angaben der Ausgabe von AWStats, die als Abbildungen folgen, erhält man

nvt = (Summe über alle Hits mit Status Codes ungleich 200, 304) + (Summe aller Robot Hits) = 25.237 + 40.757 = 65.994

und damit weniger als im Summary. Wieso Hits nicht berücksichtigt werden, kann nicht nachvollzogen werden. Die ausgeschlossenen Status Codes sind korrekt, ebenso die Angabe vt = (Summe über alle Hits aus der Host-Statistik) = 65.701.

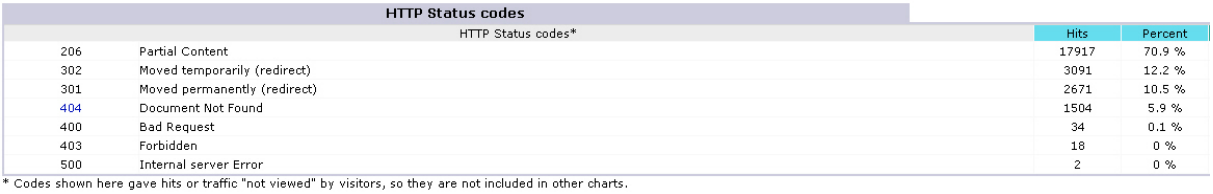


Robots/Spiders visitors (Top 10) - Full list - Last visit

47 different robots*	Hits	Bandwidth
Googlebot	16545+6	509.72 MB
Yahoo Slurp	8952+228	383.83 MB
Java (Often spam bot)	4282	124.30 MB
Python-urllib	3908	15.03 MB
mnoGoSearch search engine software	3788+4	24.54 MB
Ask	1047+14	26.89 MB
Yahoo-MMCrawler	602+1	187 Bytes
MSNBot	469+78	20.29 MB
Nagios	416	6.59 MB
ASpider (Associative Spider)	147+2	3.15 MB
Others	285+183	62.57 MB

* Robots shown here gave hits or traffic "not viewed" by visitors, so they are not included in other charts. Numbers after + are successful hits on "robots.txt" files.

Abb. 2: AWStats Robots

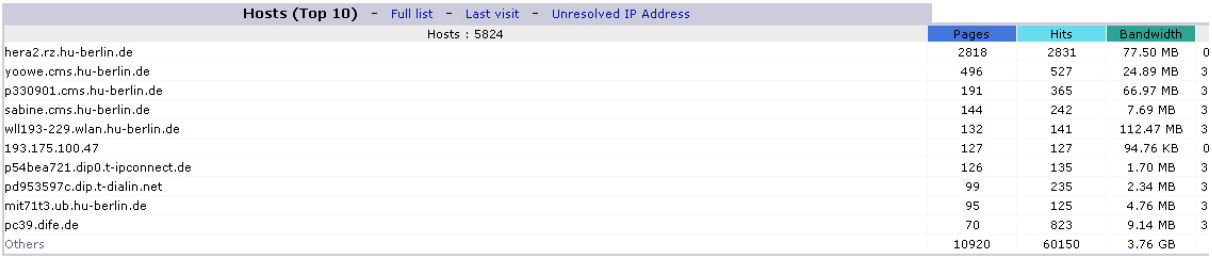


HTTP Status codes

	HTTP Status codes*	Hits	Percent
206	Partial Content	17917	70.9 %
302	Moved temporarily (redirect)	3091	12.2 %
301	Moved permanently (redirect)	2671	10.5 %
404	Document Not Found	1504	5.9 %
400	Bad Request	34	0.1 %
403	Forbidden	18	0 %
500	Internal server Error	2	0 %

* Codes shown here gave hits or traffic "not viewed" by visitors, so they are not included in other charts.

Abb. 3:AWStats Status Codes



Hosts (Top 10) - Full list - Last visit - Unresolved IP Address

Hosts : 5824

	Pages	Hits	Bandwidth	
hera2.rz.hu-berlin.de	2818	2831	77.50 MB	0
yoowe.cms.hu-berlin.de	496	527	24.89 MB	3
p330901.cms.hu-berlin.de	191	365	66.97 MB	3
sabine.cms.hu-berlin.de	144	242	7.69 MB	3
wll193-229.wlan.hu-berlin.de	132	141	112.47 MB	3
193.175.100.47	127	127	94.76 KB	0
p54bea721.dip0.t-ipconnect.de	126	135	1.70 MB	3
pd953597c.dip.t-dialin.net	99	235	2.34 MB	3
mit71t3.ub.hu-berlin.de	95	125	4.76 MB	3
pc39.dife.de	70	823	9.14 MB	3
Others	10920	60150	3.76 GB	

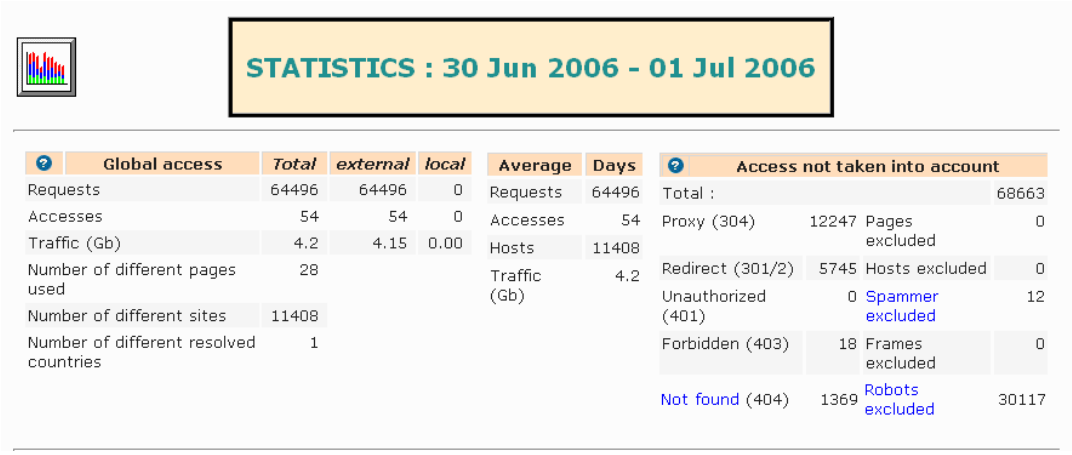
Abb. 4: AWStats Hosts

2.1.4.4 W3Perl

Tab. 6: Bewertung von W3Perl

Konfigurierbarkeit	Es kann nur angegeben werden, ob Status =206 ausgeschlossen wird, sonst ist keine Anpassung bei Status möglich, die Angabe von Hosts und Requests, die ausgeschlossen werden sollen, ist möglich.
Robots	Anzahl 27, Robot-Agents können angegeben werden, eine Liste gibt es nicht mehr.
Zählweise	Status 200, wie konfiguriert, 304 wird nicht gezählt.
Plausibilität	Nein, siehe Abbildung und erklärenden Text.
Sessionizing	Nicht vorhanden.
Ausgabe Ergebnisse	Es entstehen für die Auswertungen zum großen Teil Textfiles, aber z.B. nicht für Robots.
	• Graefe.pdf: $D_G = 3$ • seifahrt.pdf: $D_S = 30$ • Zimmer.pdf: $D_Z = 2$

Der Unterschied zwischen Access und Request ist unklar. Die Summe aus Requests, Accesses und „Access not taken into account" ist mit 133.213 kleiner als die Anzahl Zeilen des Testfiles. Es gibt noch andere Inkonsistenzen, die nicht aus dem Summary sofort zu erkennen sind, z. B. unterscheidet sich die Anzahl Robot Hits (30117) von den Hits in der Robot-Statistik (24994).



STATISTICS : 30 Jun 2006 - 01 Jul 2006

Global access	Total	external	local
Requests	64496	64496	0
Accesses	54	54	0
Traffic (Gb)	4.2	4.15	0.00
Number of different pages used	28		
Number of different sites	11408		
Number of different resolved countries	1		

Average	Days
Requests	64496
Accesses	54
Hosts	11408
Traffic (Gb)	4.2

Access not taken into account			
Total :			68663
Proxy (304)	12247	Pages excluded	0
Redirect (301/2)	5745	Hosts excluded	0
Unauthorized (401)	0	Spammer excluded	12
Forbidden (403)	18	Frames excluded	0
Not found (404)	1369	Robots excluded	30117

Abb. 5: :W3Perl Summary

2.1.4.5 Analog

Tab. 7: Bewertung von Analog

Konfigurierbarkeit	Sehr gut, Status Codes können beliebig ein- oder ausgeschlossen werden, die Angabe von Hosts und Requests, die ausgeschlossen werden sollen, ist möglich.
Robots	Die Anzahl wird nicht explizit ausgegeben, die Erkennung erfolgt nur nach Agent. Es gibt eine regelmäßig aktualisierte Liste, die allerdings in veränderter Form in das Konfigurationsfile eingefügt werden muss. Hosts mit robots.txt-Zugriff werden nicht ausgesondert.
Sessionizing	Nicht vorhanden.
Zählweise	Es werden nur Zeilen mit Status 200 und 304, wie konfiguriert, gezählt.
Plausibilität	Alle Zahlen sind nachvollziehbar.
Ausgabe Ergebnisse	Es gibt gut geeinete Ausgaben in den Formaten ascii, html, xhtml, LateX, computer (letztes Format kann beliebig angepasst werden).
Ergebnisse	• Graefe.pdf: $D_G = 4$ • seifahrt.pdf: $D_S = 34$ • Zimmer.pdf: $D_Z = 2$

Wie die folgende Abbildung zeigt, sind vollständige und nachvollziehbare Angaben im Summary enthalten.

Program started at Wed-15-Aug-2007 16:23.
Analysed requests from Fri-30-Jun-2006 01:51 to Sat-01-Jul-2006 01:38 (0.99 days).

General Summary

(**Go To**: Top: General Summary: Monthly Report: Daily Summary: Hourly Summary: Dor
Report: Search Word Report: Browser Report: Browser Summary: Operating System Re
Report: Request Report)

This report contains overall statistics.

Successful requests: 63,980
Average successful requests per day: 64,561
Successful requests for pages: 6,668
Average successful requests for pages per day: 6,727
Distinct files requested: 24,498
Distinct hosts served: 5,984
Corrupt logfile lines: 1
Unwanted logfile entries: 69,404
Data transferred: 3.84 gigabytes
Average data transferred per day: 3.88 gigabytes

Abb. 6: Analog Summary

2.1.4.6 WUMprep

Dieses Tool unterscheidet sich von den vorangegangenen dadurch, dass es wesentlich flexibler angewendet werden kann. Sein eigentlicher Zweck ist die Aufbereitung der Logfiledaten für weiteres Data Mining mit dem Tool WUM[102] (Spiliopoulou und Faulstich 1999). WUMprep ist eine Sammlung von Perl-Scripts, die nacheinander, auch in individuell festgelegter Reihenfolge, abgearbeitet werden. Die Analysemöglichkeiten sind auch aus diesem Grund sehr variabel und gehen weit über die Möglichkeiten der anderen Programme hinaus. Besonders das Sessionizing, welches nach sehr ausgefeilten Algorithmen ausgeführt wird und angepasst werden kann, hat ein viel höheres Niveau als bei den anderen Tools. Auch die Robot-Erkennung basiert auf einer Zuordnung von Sessions und bezieht Nutzerverhalten mit ein. Leider sind die Ergebnisse nicht nachvollziehbar. So werden Sessions als Robot Session erkannt, obwohl kein einziges Merkmal zutrifft.

Tab. 8: Bewertung von WUMprep

Konfigurierbarkeit	Sehr gut.
Robots	Die Anzahl wird nicht explizit ausgegeben, sondern ergibt sich aus dem Ergebnisfile. Die Liste der Robots ist von 2003. Robots werden durch Host, Agent, Zugriff auf robots.txt und Sessionizing erkannt.
Sessionizing	Ja, wird auch zur Robot-Erkennung verwendet.
Zählweise	Es werden die Zeilen mit Status 200 und 304, wie konfiguriert, gezählt.
Plausibiliät	Es gibt hier kein Summary, welches man mit dem der anderen Programme vergleichen könnte, die Ergebnisse des Sessionizings sind widersprüchlich.
Ausgabe Ergebnisse	Es gibt gut geeinete Ausgaben in den Formaten ascii und html.
Ergebnisse	• Graefe.pdf: $D_G = 3$ [103] • seifahrt.pdf: $D_S = 29$ • Zimmer.pdf: $D_Z = 0$

2.1.4.7 Deep Log Analyzer

Dieses Programm in der Lightversion ist von der Bedienung her sehr komfortabel konfigurierbar. Es können aber nur wenige Parameter konfiguriert werden. Sein größtes Handicap besteht darin, dass man nicht wählen kann, welche Zugriffe gezählt werden sollen. Die Ergebnisse unterscheiden sich sehr stark von denen der anderen Tools.

[102] WUM: Web Utilization Miner.

[103] Davon ist 1 Request aber von Mozilla/5.0 (compatible; Yahoo! Slurp; http://help, welcher ein Robot ist, aber nicht als solcher erkannt wurde.

Tab. 9: Bewertung von Deep Log Analyzer Light

Konfigurierbarkeit	Wenig Möglichkeiten, es können aber Hosts ausgeschlossen werden.
Robots	Anzahl 15, Ekennung nach Agent.
Sessionizing	Es wird wahrscheinlich zur Bildung von Visits verwendet.
Zählweise	Es kann nicht konfiguriert werden, ob und wie der Status Code berücksichtigt wird, ist nicht festzustellen.
Plausibilität	Es gibt hier kein Summary, welches überprüft werden kann.
Ausgabe Ergebnisse	Export nach html ist möglich.
Ergebnisse	• Graefe.pdf: $D_G = 583$ • seifahrt.pdf: $D_S = 171$ • Zimmer.pdf: $D_Z = 0$

Aus dem in der folgenden Abbildung gezeigten Summary können keine nachvollziehbaren Angaben entnommen werden.

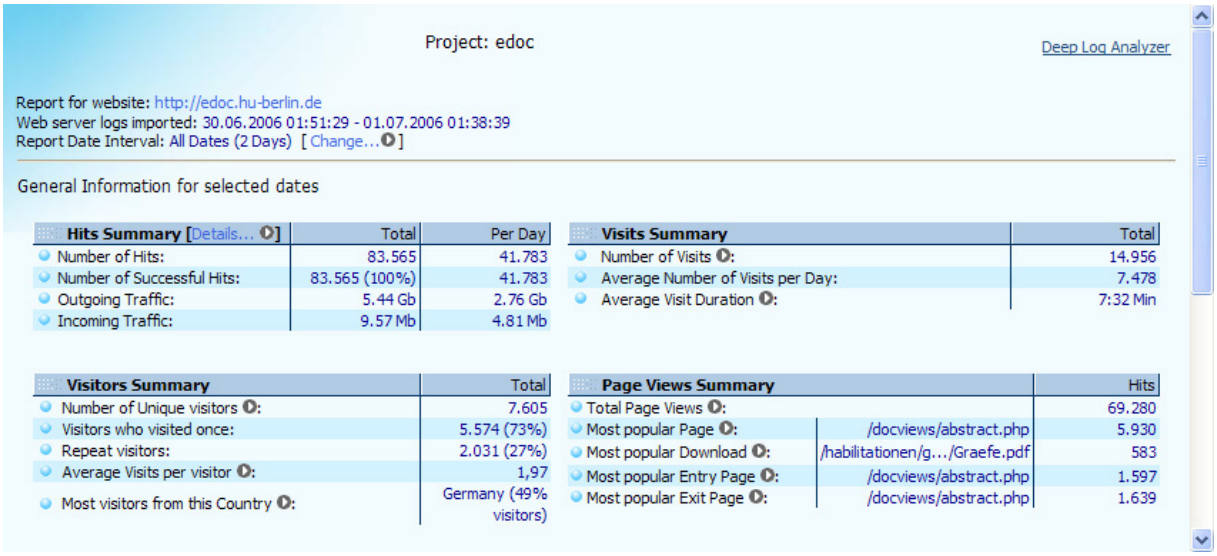


Abb. 7: Summary Deep Log Analyzer Light

Es ist unklar, welche Hits gezählt werden und wie die Angaben im Summary entstehen. Das Tool ist für die Logfileanalyse nur stark eingeschränkt brauchbar.

2.1.4.8 Ergebnisse des Vergleiches

Der Vergleich von Tools zur Logfileanalyse ist schwierig, weil die Konfigurationsmöglichkeiten unterschiedlich sind und deshalb nicht von gleichen Voraussetzungen ausgegangen werden kann. Ein Minimum an Kon-

figurationsmöglichkeiten muss jedoch vorhanden sein. Ungeeignet sind deshalb Programme, die es nicht gestatten, die Auswertung auf bestimmte Status Codes zu beschränken, da in dieser Frage dem Standard von COUNTER gefolgt werden soll, der die Zählung von Hits mit den Status Codes 200 und 304 vorsieht. Aus diesem Grund entfallen W3Perl und Deep Log Analyzer.

Da die ermittelten Anteile von Robot Hits an der Gesamtzahl der Hits zwar unterschiedlich hoch, aber in jedem Fall beträchtlich sind, muss dem Ausschluss von Robot Hits aus der Zugriffszählung besondere Aufmerksamkeit geschenkt werden. Von großer Bedeutung ist hier die Aktualität der verwendeten Listen, an Hand derer Robots identifiziert werden. Die umfangreichste Liste, die regelmäßig aktualisiert wurde, bot Analog. Im Sommer 2007 wurde die Datenbasis dafür mit der Begründung, dass solch eine Liste nur unvollständig sein kann, von der Website genommen. Man kann aber jede andere Robot-Liste verwenden, um sie in die Konfiguration von Analog einzufügen. Verwendetet man das bereits mit WUMPrep von Robots bereinigte Ausgabefile als Eingabe für AWStats, werden noch 27 verschiedene Robots erkannt, was gegen die Qualität der Roboterkennung mit WUMprep spricht. Untersuchungen haben ergeben, dass nicht die von WUMprep verwendete Robot-Liste die Ursache war. Der zweite Grund, WUMprep abzulehnen, ist das Ergebnis des Sessionizing, bei dem einige Hosts gleichzeitig einer Robot Session und einer Nicht-Robot-Session zugeordnet wurden.

Da Analog das einzige Programm ist, welches ausschließlich nachvollziehbare Aussagen liefert, muss es hier als das zuverlässigste der Auswahl angesehen werden. Dass für die ausgewählten Beispiele die Zugriffszahlen korrekt sind, kann natürlich nicht als Beweis dafür angesehen werden, dass Analog das beste Programm ist. Die Beispiele sprechen jedoch für die Auswahl von Analog. Vertrauenswürdig bei Analog ist, dass es nicht Features vorgibt, die in Wirklichkeit nicht vorhanden sind. Statt eines Sessionizing mit fehlerhaften Ergebnissen wird konsequenterweise ganz darauf verzichtet. Gegen Analog spricht, dass die letzte Version 6.0 aus dem Jarhr 2004 stammt und offenbar nicht weiterentwickelt wird.

AWStats dagegen wird weiterentwickelt. Die im Moment letzte Version 7.0 stammt aus dem Jahr 2010. Die nicht nachvollziehbaren Angaben, die auch noch in der Version 6.95 unverändert sind, sprechen nicht für dieses Tool. Allerdings kommen die Ergebnisse den Kontrollwerten nahe und die Filterung der von WUMprep nicht erkannten Robots spricht auch für die Verwendung von AWStats.

2.2 Zusammenfassung

Downloadzahlen der Publikationen von OA Repositories können mit verschiedenen Tools, die wiederum individuell konfiguriert sind, ermittelt werden. Der Vergleich der Ergebnisse für fünf verschiedene Tools gibt darüber bereits Auskunft. Selbst wenn man davon ausgehen würde, dass alle Repository-Betreiber nur AWStats oder Analog benutzen, gäbe es Unterschiede in den Ergebnissen. Die folgende Tabelle zeigt die Ergebnisse der Logfileanalyse eines Tages für drei Beispielfiles. Die Unterschiede nehmen erheblich zu, wenn die Ergebnisse für einen Monat summiert werden. Außerdem wurden Beispiele gewählt, bei denen keine gehäuften Robot-Zugriffe auftraten.

Tab. 10: Zusammenfassung der Ergebnisse für die Beispielfiles

	Graefe.pdf	**seifarth.pdf**	**Zimmer.pdf**
AWStats	0	34	0
W3Perl	3	30	2
Analog	4	34	2
WUMprep	3	29	0
Deep Web Analyzer	583	171	0

Die größten Probleme verursachen jedoch Robots. Selbst wenn alle Betreiber die gleichen Robots filtern würden, könnten damit nicht alle erfasst werden, da nicht alle Robots bekannt sind. Hinzu kommt, dass nicht alle Repositories von den gleichen Robots aufgesucht werden. Wenn das der Fall wäre, könnte man zumindest davon ausgehen, dass der Einfluss von Robots auf die Downloadzahlen überall gleich ist. Eine Möglichkeit, den Einfluss der Robots auf die Downloadzahlen zu verringern, ist die regelmäßige Beobachtung der Nutzungsstatistik. Bei der Statistik mit AWStats oder Analog kann man z. B. kontrollieren, welche Hosts die meisten Hits auslösen und diese dann, wenn es Robots sind, im Konfigurationsfile zusätzlich ausschließen. Das wäre aber eine individuelle Maßnahme, die auf die Downloadzahlen der Repositories unterschiedliche Auswirkungen hätte. Der folgenden Tabelle kann entnommen werden, wie unterschiedlich die Robot-Filterung bei AWStats und Analog erfolgt.

Tab. 11: Anteile von Robot Requests an den Hits mit Status Code 200 und 304

Programm	**Anzahl Robot Hits**	**Anteil Robot Hits in %**
AWStats	32 511	30
Analog	44 161	41

Anzahl Datensätze im Logfile mit Status Code 200 und 304: 108 141, Filterung der Robots durch die mitgelieferte Liste der Tools

Aus der Zusammenfassung wird deutlich, dass kein Tool in der Lage ist, durch Konfiguration oder Algorithmus den Ausschluss von Robotzugriffen auf zufriedenstellende Art zu lösen. Es werden viele der Möglichkeiten der Roboterkennung, wie sie in in der Liste von 1-6 aufgezählt sind, genutzt. Punkt 6 der Liste (Muster einer Session) wird aber nur unzureichend zur Erkennung von Robots angewendet. Zieht man in Betracht, dass Repository-Betreiber durch Preprocessing der Logfiles oder durch augefeiltere Filtermethoden den Aus-

schluss von Robots verbessern können[104], muss man von noch größeren Unterschieden bei der Ermittlung von Downloadzahlen ausgehen, als sie beim Vergleich von fünf Tools zu Tage treten. Downloadzahlen verschiedener OA Repositories sind nicht vergleichbar, da sie nicht mit einheitlichen Verfahren ermittelt werden und dem nur begrenzt zu kontrollierenden Einfluss von Robots ausgesetzt sind. Deshalb eignen sich Downloadzahlen, die von den IR-Betreibern individuell erhoben werden, nicht als Impact-Maß für OA-Publikationen.

2.3 Schlussfolgerungen

Die im vorangegangenen Abschnitt dargestellte Situation zu verändern, ist das Ziel der europäischen Projekte SURE[105], PIRUS[106] und OA-Statistik[107], die unter dem Dach von Knowledge Exchange (KE)[108] eng miteinander kooperieren. Inzwischen als Pseudostandard anerkannten Regelungen für die Erzeugung von Nutzungsstatistiken des bereits im Abschnitt 2.1.3 erwähnten COUNTER Code of Practice fließen in die Arbeit dieser Projekte ein und werden um Standards erweitert, die den Anforderungen von OA gerecht werden. Um vergleichbare Nutzungsstatistiken zu erzeugen, werden die Logfiles nicht mehr lokal für jedes Repository einzeln, sondern zentral analysiert. Die Downloadzahlen werden nicht nur für die Publikationen jedes Repository ermittelt, sondern werden im Fall von Dubletten summiert. Dabei soll nicht ausschließlich wie bei COUNTER, sondern auch nach anderen Standards aggregiert werden. So können für die Publikationen vergleichbare Downloadzahlen erzeugt und gleichzeitig verglichen werden, was sich bei Verwendung einer anderen Metrik ergibt. Um eine solche Infrastruktur aufzubauen, werden gemeinsame Richtlinien, die KE Usage Statistics Guidelines erarbeitet. Eine einheitliche Robot-Liste, die auf der von COUNTER veröffentlichten basiert, wird durch alle Beteiligten ergänzt.

---

[104] Bereits zitiert wurden (Bomhardt et al. 2005; Geens et al. 2006; Stassopouloua und Dikaiakos 2009), eine weitere Methode findet man in (Sponsler 2004).

[105] SURE: Statistics on the Usage of Repositories, http://www.surffoundation.nl/nl/projecten/Pages/SURE.aspx, gelesen 26.1.2011.

[106] PIRUS: Publisher and Institutional Repository Usage Statistics, http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx, gelesen 26.1.2011.

[107] Open Access Statistik, http://www.dini.de/projekte/oa-statistik/, gelesen 26.1.2011.

[108] KE ist eine gemeinsame Initiativ von Forschungsfördereinrichtungen aus Großbritannien, den Niederlanden, Dänemark und Deutschland, die den Ausbau der Informations- und Kommunikationstechnologie zum Ziel hat, http://knowledge-exchange.info/, gelesen 26.1.2011.

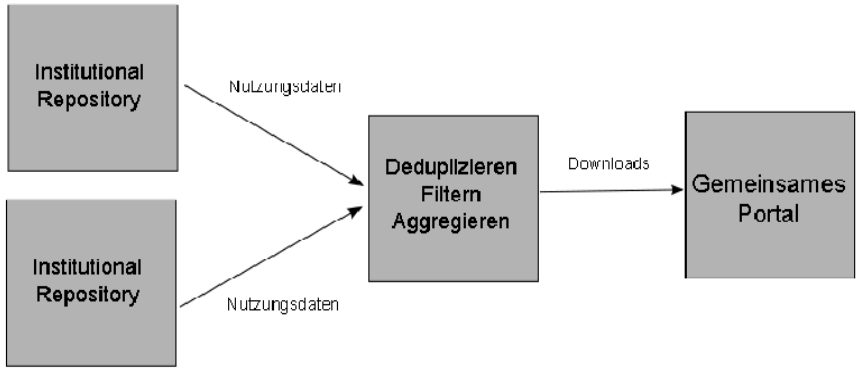


Abb. 8: Zentrale Verarbeitung von Nutzungsdaten

Diese Projekte befinden sich gegenwärtig im Teststadium und es wird noch einige Zeit vergehen, bis solch eine Infrastruktur für OA Repositories existiert. Bis dahin wird es keine Downloadzahlen für Publikationen geben, die sebst als Impact-Maße gelten oder auf deren Grundlage Impact-Maße gebildet werden können. Die Informationen aus den Downloadzahlen können aber, auch wenn die einzelne Downloadzahl nicht zuverlässig ist, in ihrer Gesamtheit dazu dienen, die Nutzung eines Repository zu analysieren.

3 Die Entwicklung der Analysemethode NoRA

3.1 Ziel und Eigenschaften

Innerhalb der OA-Bewegung erfüllen IR wichtige Funktionen. Sie dienen den Mitarbeitern einer Institution zum Veröffentlichen von Pre- und Postprints nach dem OA-Prinzip und sind dadurch ein Teil der „Green Road to Open Access“. Autoren und Herausgeber nutzen zunehmend IR, um Zeitschriften und andere fortlaufende Sammelwerke, Konferenzproceedinges und andere Publikationsformen von der Papierform in die elektronische Form zu überführen oder von Beginn an nur online zu veröffentlichen. Dadurch sind IR bereits jetzt ein fester Bestandteil der „Golden Road to Open Access“. Zur Unterstützung von Herausgebern bei der Qualitätssicherung werden Begutachtungssysteme angeboten. Zu diesen für die OA-Bewegung spezifischen Funktionen kommen Aufgaben, die sich aus den Anforderungen der Institution ergeben, z. B. die Veröffentlichung von Hochschulschriften und die Bereitstellung von Arbeitsmaterialien, wie es für Universitäten typisch ist. Das Spektrum der angebotenen Dienste ist breit und erfordert hohen personellen und finanziellen Aufwand. Die Ressourcen sind begrenzt und reichen gerade aus, die täglich anfallenden Arbeiten wie die inhaltliche Erschließung von Publikationen, die Pflege des Systems oder die Beratung von Autoren und Herausgebern zu leisten. Dabei gerät im Routinebetrieb oft aus dem Blick, dass die Weiterentwicklung des IR als Ganzes nicht vernachlässigt werden darf. Ein einmal erhaltenes DINI-Zertifikat sollte zwar garantieren, dass Qualitätskriterien formal erfüllt werden, den graduellen Unterschieden wird jedoch wenig Beachtung geschenkt. Ein zentrales Kriterium, bei dem solche graduellen Unterschiede wichtig sind, ist die Sichtbarkeit der Publikationen im Internet als Voraussetzung für deren Nutzung und barrierefreien Verbreitung des in ihnen enthaltenen Wissens als wichtigstem Ziel der OA-Bewegung.

Um die Akzeptanz von OA zu steigern, muss eine Strategie verfolgt werden, welche die Steigerung der Sichtbarkeit der OA-Publikationen zum Ziel hat. Das wirksamste Mittel dazu ist die Akquisition von Publikationen unter drei Gesichtspunkten:

1. Erhöhung der Gesamtzahl der Publikationen
2. Erhöhung der Anzahl von Publikationen, die eine hohe Nutzungshäufigkeit erwarten lassen
3. Verbesserung der Sichtbarkeit von Publikationen mit geringer Nutzung

Potentielle Nutzer von IR werden diese erst dann in ihre Recherchen einbeziehen, wenn die Anzahl an Publikationen eine kritische Masse überschritten hat. Mit der Schaffung eines Netzwerks von OA Repositories und eines gemeinsamen Rechercheportals soll gerade solch eine kritische Masse erreicht und überschritten werden. Dieses Netzwerk existiert in Ansätzen und ist auf die Publikationen der beisteuernden IR angewiesen. Für ein IR geht es aber auch um die Stärkung der Akzeptanz an der Institution selbst und von dieser Perspektive aus müssen die genannten Punkte ebenso Beachtung finden. Während die Entwicklung der Anzahl einfach an den Metadaten ablesbar ist, erfordert die Beachtung des 2. und 3. Punktes die Analyse und Auswertung von Nutzungsdaten. Es muss festgestellt werden, ob es Gruppen von Publikationen gibt, die sich in ihrer Nutzungshäufigkeit signifikant voneinander unterscheiden, um Maßnahmen wie die gezielte Akquisition von

Publikationen einer formalen oder inhaltlichen Gruppe oder die Ermittlung der Ursachen vergleichsweise schlechter Sichtbarkeit einleiten zu können. Damit ist das Ziel der Analysemethode NoRA umschrieben. Sie wurde so konzipiert, um sie vor allem für IR von Universitäten einzusetzen, ist aber flexibel genug, um sie für jede Art von OA Repository anwenden zu können.

Wie aus dem vorangegangenen Kapitel hervorgeht, muss bei der Analyse und Auswertung von Nutzungsdaten berücksichtigt werden, wie sie erhoben wurden und welchen Einflüssen sie unterliegen. Es ist keinesfalls möglich, Nutzungszahlen etwa durch Bildung von Mittelwerten über Kategorien von Merkmalen, die sich aus den Metadaten der Publikationen ergeben, zu analysieren. Das wäre zwar einfach, aber irreführend. Dem Charakter des Datenmaterials muss Rechnung getragen werden, indem adäquate Methoden der Mathematischen Statistik zur Anwendung kommen. Trotzdem soll die Analyse durch die Betreiber selbst durchführbar sein, ohne dass spezielle Kenntnisse der Mathematischen Statistik vorausgesetzt werden müssen. Das erfordert eine Auswahl einfach anzuwendender Verfahren, die trotzdem das Notwendige leisten. Die dabei benutzten Software-Tools müssen an den Institutionen verbreitet sein und ohne zusätzliche Kosten bereitgestellt werden können. Ihre Verwendung darf keinen hohen Einarbeitungsaufwand erfordern. An Vorkenntnissen werden lediglich solche vorausgesetzt, wie sie im Umgang mit Datenmengen, die in Art und Umfang typisch für OA Repositories sind, vorhanden sein müssen. Die Darstellung von NoRA ist so gewählt, dass sie in nacheinander abzuarbeitende Schritte eingeteilt und am Beispiel nachvollziehbar ist. Das Beispiel orientiert sich an inhaltlichen und formalen Merkmalen, wie sie in den Metadaten für wissenschaftliche Publikationen von universitären IR vorhanden sind. Die Analysemethode NoRA kann aber auf jede Art von Merkmalen angewendet werden und ist deshalb flexibel einsetzbar.

Nutzungsdaten von elektronischen Publikationen werden seit Mitte der 1990er Jahre untersucht. Schon damals galt das Interesse dem Zusammenhang zwischen der Nutzung von Artikeln einer Online-Zeitschrift und dem späteren Citation Impact. Das Hauptaugenmerk galt aber zunächst vor allem der Nutzung digitaler Bibliotheken[109]. Von 2000 an verstärkte sich das Interesse an der Analyse von Nutzungsdaten elektronischer Publikationen und es wurde versucht, das Konzept des Citation Impact auf einen Download Impact zu übertragen.[110] Seit dem Aufschwung der OA-Bewegung im Jahr 2001 wird der Einfluss von OA auf den Citation Impact untersucht. Das am meisten untersuchte OA Repository ist arXiv.org (siehe Kapitel 1). Die Untersuchung von Nutzungsdaten eines deutschen IR ist bisher nicht bekannt. Dafür gibt es vermutlich zwei Gründe. Die Anzahl der auf deutschen IR veröffentlichten Publikationen, die zugleich im Science Citation Index erfasst sind, ist sicher zu gering, um einen Einfluss der Nutzung überprüfen zu können. Zum anderen ist klar, dass die Nutzungsdaten, wie sie im Moment erfasst werden, mit vielen Mängeln behaftet sind und der Vergleich der Nutzung verschiedenen IR nicht möglich ist. Das Hauptinteresse der Forschung im Zusammen-

[109] In der Arbeit von Kaplan und Nelson (Kaplan und Nelson 2000) findet man eine Zusammenstellung von Artikeln bis zum Jahr 2000, die sich mit der Analyse von Nutzungsdaten beschäftigen.

[110] Erwähnt wurden bereits im Abschnitt 1.2.4 (Bollen et al. 2003; Bollen und van de Sompel 2008; Brody et al. 2006).

hang mit Nutzungsdaten von OA Repositories ist deshalb sowohl in Deutschland als auch international auf eine einheitliche Erhebung und standardisierte Aggregation der Daten gerichtet. Wie und zu welchem Zweck Nutzungsdaten von IR trotz dieser Mängel sinnvoll verwendet werden können, wurde bisher nicht untersucht. Insofern stellt die Analysemethode NoRA ein Novum dar.

3.2 Vorgehensweise

Bei der Wahl der zu verwendenden Software wurden Gesichtspunkte berücksichtigt, die sich aus der Art des Datenmaterials und der im vorangegangenen Abschnitt aufgeführten gewünschten Eigenschaften der Analysemethode NoRA ergeben. Wie schon mehrfach bemerkt, ist es aufgrund der Erkenntnisse über die Erhebung von Nutzungsdaten und die sie beeinflussenden Faktoren unumgänglich, Verfahren der Mathematischen Statistik anzuwenden. Gleichzeitig benötigt man ein Instrument für das Datenmanagement, d. h. ein Programm zur Aufbereitung der Daten, so dass die Methoden angewendet und die gestellten Fragen beantwortet werden können. Des Weiteren ist es wichtig, die Ergebnisse ohne großen Aufwand grafisch darzustellen. Ein nicht zu vernachlässigender Gesichtspunkt für die Wahl der Software ist einfache Bedienbarkeit, ohne dass Spezialkenntnisse aus dem Bereich der Mathematischen Statistik erforderlich sind. Die Wahl fiel auf deshalb auf SPSS[111], welches alle diese Aspekte berücksichtigt und außerdem den Vorteil hat, in der Regel an der Institution vorhanden zu sein. Alle Untersuchungen wurden mit SPSS durchgeführt.

Für den edoc-Server, dem IR der Humboldt-Universität, werden seit Jahren monatliche Downloadzahlen, ab sofort immer kurz Downloads genannt, berechnet und zusammen mit den Metadaten der Publikationen in einer Datenbank gespeichert. Daraus wurden zunächst die bis Dezember 2007 vorhandenen Metadaten und die zugehörigen Nutzungsdaten des Jahres 2007 extrahiert. Diese Daten bildeten das Ausgangsmaterial, anhand dessen die charakteristischen Eigenschaften der Nutzungsdaten untersucht wurden. Das Ergebnis war, dass trotz der deutlich sichtbaren Schwankungen der Werte Strukturen erkennbar sind und es grundsätzlich möglich ist, sinnvolle Analysen von monatlichen Nutzungsdaten durchzuführen. Daraufhin wurde überprüft, welche Metadaten in die Analyse einbezogen werden können, um für den Betreiber relevante Fragestellungen zu bearbeiten. Hier zeigte sich, dass anhand der Metadaten, die Angaben zum formalen Publikationstyp und eine inhaltliche Klassifikation enthalten, Gruppen von Publikationen gebildet werden können, deren Downloads sich signifikant unterscheiden. Es erwies sich weiterhin, dass die genaue Zeitangabe der Online-Veröffentlichung benötigt wird, um die Entwicklung des Bestandes zu analysieren. Des Weiteren wurde ermittelt, wie groß die Fallzahl in Gruppen, die sich aus den Metadaten ergeben, mindestens sein muss, damit belastbare Aussagen gemacht werden können und welchen Einfluss Zeit und Alter der Publikationen haben. In der Folge wurde die Datenbasis um Metadaten und Nutzungsdaten bis März 2010 erweitert. Alle geeignet

[111] SPSS bedeutet usprünglich “Statistical Package for the Social Sciences” (http://www.spss.com/de/, gelesen 21.1.2011). Das Programm liegt aktuell als IBM SPSS Statistics 19 vor. Die Vorgängerversion, mit der hauptsächlich gearbeitet wurde, trug den Namen PASW (Predictive Analysis SoftWare).

erscheinenden statistischen Verfahren wurden auch an diesen Daten erprobt. Das letztendlich ausgewählte Verfahren von Kruskal-Wallis bildet den Kern der Analysemethode NORA.

Nachdem die Fragen der grundsätzlichen Auswertbarkeit monatlicher Nutzungszahlen und der mindestens erforderlichen Metadaten geklärt waren, wurde versucht, weitere Daten zu akquirieren, um die Anwendbarkeit von NoRA in einer Studie zu erproben. Es wurden mehrere Betreiber von IR direkt angesprochen. Meist bedurfte es eines kurzen Schriftwechsels, um das Anliegen zu verdeutlichen und Daten zu erhalten, die sich für die Anwendung von NoRA eignen. Die Metadaten sollten neben dem Datum der Online-Veröffentlichung mindestens eine formale und inhaltliche Klassifikation und wenn möglich zusätzlich die Zuordnung zu einer Institution wie beispielsweise einer Fakultät enthalten. Die Nutzungsdaten sollten monatlich aggregiert und für Volltexte erhoben worden sein. Im Laufe dieser Kommunikation entstand ein Anschreiben für bisher noch nicht persönlich kontaktierte IR-Betreiber, in dem das Vorhaben erläutert wird und ein Datenblatt mit der Beschreibung der benötigten Daten (siehe Anhang A). Dieses Schreiben wurde an weitere IR-Betreiber weitergeleitet.

Von den IR werden Nutzungsdaten in unterschiedlicher Form angeboten. Um für die Studie Daten bereitzustellen, müssen sie in der Regel von den Betreibern in ein Format transformiert werden, welches für die Auswertung geeignet ist. Das ist sicherlich ein Grund, dass nur weitere drei Institutionen Daten zur Verfügung stellten. Als anderer Grund wurde genannt, dass derart strukturierte Daten nicht vorhanden sind. Insgesamt liegen nun Daten von vier deutschen universitären IR vor:

- edoc (edoc-Server der Humboldt-Universität zu Berlin)[112]
- Elektronische Hochschulschriften der Universität Stuttgart (ehsStu)[113]
- HeiDOK (Der Heidelberger Dokumentenserver, Universitätsbibliothek Heidelberg)[114]
- SciDok (Der Wissenschaftsserver der Universität des Saarlandes)[115]

Bei allen vier Beteiligten handelt es sich um deutsche universitäre IR, die einige spezifische Eigenschaften aufweisen. Z. B. bilden Qualifikationsarbeiten wie Diplom-, Magister- und Bachelorarbeiten und Dissertationen einen großen Anteil an den Publikationen. Die IR HeiDOK, SciDok und ehsStu basieren auf der Software OPUS[116], die Software von edoc ist eine Eigenentwicklung der Humboldt-Universität[117]. Die Nut-

[112] http://edoc.hu-berlin.de/, gelesen 21.1.2011.

[113]Die Kurzbezeichnung ehsStu wird nicht von der Universität Stuttgart, sondern nur in dieser Arbeit verwendet. http://elib.uni-stuttgart.de/opus/, gelesen 21.1.2011.

[114] http://archiv.ub.uni-heidelberg.de/volltextserver/, gelesen 21.1.2011.

[115] http://scidok.sulb.uni-saarland.de/, gelesen 21.1.2011.

[116] OPUS ist eine Open Source-Software unter der GNU General Public License für den Betrieb von institutionellen Dokumentenservern bzw. Repositorien. OPUS steht für Online Publikationsverbund Universität Stuttgart und wurde dort Ende der 1990er Jahre vom Rechenzentrum der Universitätsbibliothek entwickelt. Seitdem wird OPUS mit nationalen

zungsdaten wurden mit AWStats und Analog erzeugt. Da sich die Metadaten in vielen Punkten unterschieden, wurden Kategorien von formalen und inhaltlichen Merkmalen gebildet, in die sich die Daten aller Studienteilnehmer einordnen lassen. Die ursprünglich geplante und am Beispiel edoc getestete Auswertung von Nutzungsdaten von Frontdoor, PDF- und HTML-Volltexten musste auf die Auswertung von PDF-Files reduziert werden. Die anhand der Daten von edoc entwickelte Analysemethode wurde auf die gemeinsamen Kategorien und die Nutzungsdaten aller IR angewendet und konnte dadurch erfolgreich erprobt werden. Gemeinsame Kategorien wurden gebildet, um nachzuweisen, dass die Anwendung nicht nur unter speziellen Bedingungen Ergebnisse bringt und dass bei allen vier IR signifikante Downloadunterschiede zwischen Kategorien oder Gruppen von Kategorien vorhanden sind, die Aussagen über die generelle Nutzung von IR erlauben. Es wird gezeigt, wie die Analysemethode auch auf spezielle Anwendungsfälle angepasst werden kann. Die Auswahlkriterien für die zu analysierenden Daten wurden zunächst anhand von edoc aufgestellt und begründet. In der Folge stellte sich heraus, dass diese Kriterien auch für die Daten von HeiDOK und SciDok anwendbar sind.

Zum Schluss wurden alle notwendigen Tätigkeiten von der Datenaufbereitung über die Anwendung statistischer Verfahren bis zur Darstellung der Ergebnisse systematisiert, so weit wie möglich formalisiert und die Analysemethode NoRA in Form von 6 nacheinander auszuführenden Arbeitsschritten dargestellt.

3.3 Das Datenmaterial

Die folgende Tabelle gibt einen Überblick über alle verwendeten Daten. Es handelt sich ausnahmslos um Daten von IR mit DINI-Zertifikat.

Tab. 12: Überblick über die Daten der vier beteiligten Institutional Repositories

	edoc (Berlin)	**ehsStu (Stuttgart)**	**HeiDOK (Heidelberg)**	**SciDok (Saarbrücken)**
DINI-Zertifikat[118]	2007	2004	2007	2004
Publikationen bis	3/2010	11/2008	7/2009 12/2010	6/2009
Inhaltliche Klassifikation	DDC	DDC	DDC	DDC
Formale Klassifikation	edoc-Types	OPUS-Types	OPUS-Types	OPUS-Types
Zusätzliche Institution	Fakultät	Fakultät	Fakultät	Fakultät
Nutzungsdaten	1/2007-3/2010	1/2007 – 12/2008*	2/2009 bis 7/2009 1/2010 bis 12/2010	1/2007 bis 6/2009

* 2x halbjährlich 2007, ein Wert 2008 (11-12)

Partnern kooperativ weiterentwickelt (http://www.kobv.de/opus4/ueberblick/, gelesen 2.4.2011). Die vorliegenden Daten entstammen dem Vorgänger OPUS 3.

[117] Siehe http://edoc.hu-berlin.de/e_info/dokumentation.php, gelesen 2.4.2011.

[118] Das angegebene Zertifikat ist das zum Zeitpunkt der Datenerhebung gültige.

Aus Heidelberg wurden kurz vor Beendigung der Arbeiten weitere Daten bereitgestellt. Es handelt sich dabei um Metadaten bis 12/2010 und monatliche Nutzungsdaten für 2010. Diese Daten wurden nicht in die Erarbeitung der Methode einbezogen. Die Methode konnte aber erfolgreich auf diese Daten angewendet werden. Die Ergebnisse der Analyse dieser neueren Nutzungsdaten werden zusammen mit den anderen Ergebnissen dargestellt. Bei allen vier IR ist das Datum der Online-Veröffentlichung in den Metadaten tagesgenau enthalten. In den Nutzungsdaten von ehsStu, SciDok und HeiDOK sind Downloads von PDF-Files enthalten.

- **edoc (Berlin):** Es liegen monatliche Downloads von 1/2007 bis 3/2010 vor. Darin sind Daten für die Frontdoor und die verschiedenen Volltextversionen PDF und HTML enthalten. Die Downloads für HTML-Volltexte, die in der Regel aus mehreren Teilfiles bestehen, sind das Maximum der Downloads der einzelnen Files.
- **ehsStu (Stuttgart):** Es liegen halbjährliche Downloads für 1-6/2007 und 7-12/2007 und Downloads für 11-12/2008 von Volltexten vor. Um die Daten nutzen zu können, wurden die zweimonatigen Downloads mit 3 multipliziert. Es wurden keine Angaben gemacht, um welches Volltextformat es sich handelt, es sind aber nur PDF-Files bekannt. Gibt es zu einer Publikation mehrere Volltexte, werden die Downloads der einzelnen Volltexte summiert angegeben.
- **HeiDOK (Heidelberg):** Es liegen monatliche Downloads von 02/2009 bis 07/2009 und 01/2010 bis 12/2010 für Volltexte vor. Es wurden keine Angaben gemacht, um welches Volltextformat es sich handelt, es sind aber nur PDF-Files bekannt. Gibt es zu einer Publikation mehrere Volltexte, werden die Downloads summiert angegeben.
- **SciDok (Saarbrücken):** Es liegen monatliche Downloads von 1/2007 bis 6/2009 für Volltexte vor. Es wurden keine Angaben gemacht, um welches Volltextformat es sich handelt, es sind aber nur PDF-Files bekannt. Für Publikationen mit mehrere Volltexten wurden Downloads für die Teilfiles angegeben. Um Vergleichbarkeit herzustellen, wurden die Downloads der Teilfiles summiert.

Nachdem sich herausgestellt hatte, dass es nur bei edoc Downloads für HTML-Files gibt, wurde sich ausschließlich auf das Format PDF konzentriert. Als Downloads wurden auch bei edoc nur noch die Zahlen für die PDF-Files verwendet. Die Analyse der Downloads der Frontdoors wurde ebenfalls verworfen, da solche Nutzungszahlen nur für edoc und HeiDOK vorlagen. Downloads von Publikationen, für die keine Volltexte vorhanden sind, das sind z. B. Abstracts oder audiovisuelle Publikationen, werden nicht ausgewertet.

Erwartungsgemäß gibt es bei den vier IR Unterschiede in der Erfassung der Metadaten. Das betrifft vor allem die formale Klassifikation, aber auch bei der inhaltlichen Klassifikation sind Abweichungen zu berücksichtigen. Das genaue Vorgehen bei der Bildung von gemeinsamen Kategorien der Klassifikationsmerkmale wird in den folgenden Abschnitten beschrieben.

3.3.1 Formale Klassifikation

Die formale Klassifikation erfolgt nach formalen Publikationstypen. Zu erwarten war, dass bei edoc teilweise andere Kategorien von Publikationstypen verwendet werden als bei den OPUS-basierten IR. Bei genauerer Prüfung ergaben sich aber auch Unterschiede innerhalb der OPUS-basierten Systeme. Um eine einheitliche Anwendung von NoRA erproben zu können, wurden gemeinsame Kategorien des formalen Publikationstyps gebildet, was zu einer Vergröberung und dadurch auch zu Informationsverlust führt. Andererseits ist eine Beibehaltung der Vielzahl von Kategorien nicht möglich, da das zu einer Verringerung der Fallzahl in den Gruppen führt, was die Analyse erschwert[119].

Im Folgenden wird die formale Klassifikation der Publikationen als Typ bezeichnet. Bei den Daten von HeiDOK, SciDok und ehsStu wird der Typ durch eine OPUS-Type-ID oder durch eine verbale Bezeichnung, das Label, codiert. Es gab einige Unterschiede formaler, aber auch inhaltlicher Art. Die folgende Tabelle enthält eine Übersicht über die Verwendung der Typen.

Tab. 13: Übersicht über die Verwendung von OPUS-Typen in den vorliegenden Daten

ID	HeiDOK (Heidelberg)	ehsStu (Stuttgart)	SciDok (Saarbrücken)
1		Anleitung (Manual)	Anleitung (Manual)
2	Aufsatz	Aufsatz	Aufsatz
4	Buch (Monographie)	Buch (Monographie)	Buch (Mongraphie)
5		InBuch (Kapitel / Teil einer Monographie)	InBuch (Kaptel / Teil einer Monographie)
7	Abschlussarbeit (Bachelor, Master, Diplom, Magister etc.)	Diplomarbeit, Magisterarbeit	Diplomarbeit, Magisterarbeit, Staatsexamensarbeit
8	Dissertation	Dissertation	Dissertation
9		Festschrift	Festschrift
11	Journal (Komplette Ausgabe eines Zeitschriftenheftes)		Journal (Komplette Ausgabe eines Zeitschriftenheftes)
15	Proceedings (Komplette Ausgabe einer Konferenz etc.)	Proceedings (Komplette Ausgabe einer Konferenz etc.)	Proceedings (Komplette Ausgabe einer Konferenz etc.)
16		InProceedings (Aufsatz / Paper einer Konferenz etc.)	InProceedings (Aufsatz / Paper einer Konferenz etc.)
17	Forschungsarbeit (Research Paper)	ResearchPaper	ResearchPaper
19		Studienarbeit	
20	Report (Bericht)	Report (Bericht)	Report (Bericht)
21	Video		
22	Preprint (Vorabdruck)	Preprint (Vorabdruck)	Preprint (Vorabdruck)
23	Verschiedenes (u.a. Rezensionen)	Sonstiges	Sonstiges
24	Habilitation	Habilitation	Habilitation
25			Bachelor Thesis

[119] Darauf wird später ausführlicher eingegangen.

26			Vorlesung
94	Audiodatei		Audiodatei
95	Dissertationsabstract		Dissertationsabstract
98	Abschlussarbeit - Abstract		Abschlussarbeit - Abstract

gelb: Der Typ tritt bei allen 3 IR auf, hat aber unterschiedliche Labels
grau: Der Typ ist nicht vorhanden
grün: Der Typ wird auch anders codiert

Bei OPUS werden Bezeichnungen vorgegeben, die zwar verändert werden können, aber eindeutig auf den gleichen Typ bezogen sind. Stichproben ergaben, dass die Zuordnung der Publikationen zu den Typen oft verschieden ist und nicht alle Möglichkeiten ausschöpfen, z. B. wird bei HeiDOK nicht zwischen Manual und Buch unterschieden, sondern ein Handbuch als Buch eingeordnet oder in Saarbrücken ein Handbuch unter „Sonstiges“. Unter „Sonstiges“ finden sich sehr viele Publikationen, die eindeutig anders einzuordnen gewesen wären. Lediglich in Stuttgart werden Publikationen wie Prüfungsordnungen, Fragebögen, Studienführer usw. konsequent als „Sonstiges“ verstanden. Die Anzahl der Kategorien verleitet offenbar dazu, ungenaue Klassifikationen durchzuführen. Ausserdem überschneiden sich Kategorien wie im Fall „ResearchPaper“, deren Publikationen ebenso gut in eine andere Kategorie einzuordnen wären. Um eine möglichst korrekte Klassifikation vornehmen zu können, muss bereits nur unter Beachtung von OPUS eine Vergröberung der Kategorien erfolgen. Aus Saarbrücken wurde eine Tabelle mit der Zuordnung von englischen Typbezeichnungen zu deutschen Typbezeichnungen zur Verfügung gestellt, aus der vermutet werden kann, dass die Empfehlung zur Gestaltung der OAI-Schnittstelle für OPUS-Publikationstypen umgesetzt wurde (DINI 2005).

Bei edoc gibt es eigene selbstdefinierte Publikationstypen, die sich von den OPUS-Typen in der Bezeichnung unterscheiden, aber teilweise auf den gleichen Empfehlungen für die Gestaltung der OAI-Schnittstelle beruhen.

Tab. 14: Formale Publikationstypen von Volltexten bei edoc

Artikel einer Zeitschrift
Zeitschriftenausgabe
Buch-Kapitel oder Aufsatz
Buch
Konferenzbeitrag
Konferenzband
Beitrag in Sammelband
Sammelband
Band einer Schriftenreihe
Master
Dissertation
Historische Dissertation
Habilitation
Report
Öffentliche Vorlesung.

Aus den vorangegangenen Tabellen und dem Ergebnis der Stichproben bei den verschiedenen IR muss man den Schluss ziehen, dass es Unterschiede bei der Interpretation der Kategorien des Typs gibt, selbst wenn es sich um das gleiche OPUS-System handelt. Sogar innerhalb eines IR wurden die Typbezeichnungen unterschiedlich interpretiert. Ein weiteres Ergebnis der Stichproben ist, dass es sich nicht nur um generell verschiedene Interpretationen handelt, sondern die Zuordnungen inkonsistent erfolgten, wie das Beispiel Buch/Handbuch zeigt. Die Konsequenz für die Definition des Merkmals Formaler Publikationstyp, dem die Daten aller vier IR zugeordnet werden können, ist eine Verminderung der Anzahl der Kategorien, bei der die Praxis der Zuordnung bei den einzelnen IR berücksichtigt wird.

Eindeutig zuzuordnen sind:

- Dissertation **(A)**
- Habilitation **(B)**
- Abschlussarbeit ab Bachelor **(C)**
- Festschrift (entspricht der öffentlichen Vorlesung bei edoc) **(D)**

Zusammenzufassen zu Einzelpublikation (E):

OPUS-Typ	edoc-Typ
• Aufsatz	• Artikel einer Zeitschrift
• InBuch (Kapitel / Teil einer Monographie)	• Buch-Kapitel oder Aufsatz
• InProceedings (Aufsatz / Paper einer Konferenz etc.)	• Beitrag in Sammelband
• ResearchPaper (Forschungsarbeit)	• Band einer Schriftenreihe
• Report (Bericht)	• Konferenzbeitrag
• Preprint (Vorabdruck)	• Report

Unter dem Begriff Einzelpublikation werden zusammengefasst:

- alle unselbständigen Publikationen, die definitiv Bestandteil einer selbständigen Publikation sind, aber nicht zu D gehören
- Publikationen, die nach Informationsgehalt und Umfang den zuerst genannten unselbständigen Publikationen vergleichbar sind

Zu den Einzelpublikationen zählen damit auch geringe Werke und graue Literatur, wenn es sich dabei um wissenschaftliche Inhalte handelt.

Zusammenzufassen zu Komplettpublikation (F):

OPUS-Typ	edoc-Typ
• Anleitung (Manual) • Buch (Monographie) • Journal (Komplette Ausgabe eines Zeitschriftenheftes) • Proceedings (Komplette Ausgabe einer Konferenz etc.)	• Zeitschriftenausgabe • Konferenzband • Sammelband • Buch

Unter dem Begriff Komplettpublikation werden selbständige Publikationen zusammengefasst, zu denen ein kompletter Volltext aller Bestandteile existiert und die nicht zu A, B, C, oder D gehören.

Keine Kategorie existiert für

OPUS-Typ	edoc-Typ
• Da kein Volltext Video Audiodatei Dissertationsabstract Abschlussarbeit Abstract • Sonstiges, weil zu inhomogen • Da nur bei einem IR vorhanden und keinem anderem Typ zugeordnet werden kann Vorlesung Studienarbeit [120]	Historische Dissertationen, weil dieser Publikationstyp nur bei einem IR vorkommt und keinem anderen Typ zugeordnet werden kann. Es existieren außerdem keine PDF-Files davon.

Publikationen vom Typ „Sonstiges“ werden in der Analyse berücksichtigt, der Typ wird jedoch als fehlender Wert behandelt. Die Publikationen mit den restlichen Typen der Tabelle, das sind alle Publikationen, für die es keine Volltexte gibt, werden aus der Analyse völlig ausgeschlossen.

Auf diese Weise wird die Einteilung der Publikationen in formale Publikationstypen stark vereinfacht, was notwendig ist, um die Publikationen aller vier IR einordnen zu können. Andererseits wird sie auch korrekter, da keine Einteilung mehr suggeriert wird, die nicht eingehalten wird. Es ist natürlich nicht auszuschließen, dass auch in der vergröberten Form Fehler in der Zuordnung vorhanden sind.

3.3.2 Inhaltliche Klassifikation

Bei allen vier IR gibt es eine inhaltliche Klassifikation der Publikationen nach den hundert Klassen der zweiten Ebene der DDC. Bei edoc können Zuordnungen zu mehreren Klassen vorhanden sein, bei den übrigen IR ist immer nur eine Klasse angegeben.

Tab. 15: Inhaltliche Klassifikationen

	DDC	Struktur
edoc (Berlin)	Pro Publikation können mehrere DDC angegeben werden. Die Vollständigkeit ist für die einzelnen Publikationstypen unterschiedlich, erfasst wurde 3-stellig.	Für Dissertationen ist die Fakultät immer vorhanden, sonst unterschiedlich vollständig.
ehsStu (Stuttgart)	DDC einmal pro Publikation vorhanden und konsequent 3-stellig erfasst.	Für Dissertationen ist die Fakultät immer vorhanden, sonst unterschiedlich vollständig.
HeiDOK (Heidelberg)	DDC einmal pro Publikation vorhanden und bis auf wenige Ausnahmen 3-stellig erfasst.	Angabe einer Institution für alle Publikationen.
SciDok (Saarbrücken)	DDC einmal pro Publikation vorhanden und bis auf wenige Ausnahmen 3-stellig erfasst.	Fachgebiete sind in Form von Struktureinheiten der Universität angegeben.

Während in den Metadaten von ehsStu, HeiDOK und SciDok bis auf wenige Ausnahmen für jede Publikation ein Wert DDC erfasst wurde, ist die Erfassung von DDC bei edoc nicht vollständig.

Wie bei der formalen Klassifikation ist es das Ziel, inhaltliche Kategorien zu bilden, in die alle Publikationen eingeordnet werden können. Zuerst wurde untersucht, ob man die DDC dazu benutzen kann, Rückschlüsse auf Struktureinheiten zu ziehen und die Daten nach solchen Struktureinheiten getrennt zu analysieren, und so eine Klassifikation nach bestimmten Fachgebieten, wie sie sich auch in der Bezeichnung der Struktureinheiten widerspiegeln, durchzuführen. Der Vorteil davon wäre, dass man mit Struktureinheiten Herausgeber und Autoren gezielter adressieren könnte, als das durch DDC-Angaben möglich ist. In deutschen Universitäten sind die Fakultäten Lehr- und Verwaltungseinheiten mit unterschiedlichen Zuordnungen und Zusammenfassungen von Fachgebieten. Neben den Fakultäten mit mehreren Instituten gibt es die Verwaltung, zentrale Einrichtungen und Einrichtungen, die Fakultäten zugeordnet sind. Alle diese universitären Struktureinheiten sollen hier als Institution bezeichnet werden. Anhand der von vier universitären IR vorliegenden Daten wurde untersucht, ob man aus der DDC mit einer zu definierenden Wahrscheinlichkeit auf Institutionen schließen kann, z. B. aus DDC=340 auf die Juristische Fakultät und aus DDC=330 auf die Wirtschaftswissenschaftliche Fakultät. In den Metadaten von Dissertationen aus ehsStu (Stuttgart) und edoc (Berlin) sind Fakultäten und damit indirekte Zuordnungen von DDC zu Fakultäten enthalten, die aber nicht eindeutig sind. In den Metadaten von HeiDOK und SciDok sind zwar andere Institutionen als Fakultäten enthalten, die allerdings nicht standardisiert erfasst sind. Dadurch sind Zuordnungen von DDC zu den Institutionen schwierig oder unmöglich.

Mit heuristischen Methoden wurde versucht, eine Klassifikation zu entwickeln, welche die Information DDC mit der Information „Institution“ kombiniert, um eine Einteilung in fachliche Kategorien zu erhalten, die grob genug ist, auf möglichst viele oder sogar alle deutschen universitären IR angewendet zu werden, andererseits aber Auskunft geben kann, aus welcher Institution innerhalb der Universität die Publikationen stammen. Eine solche Zuordnung ist für eine einzelne Universität möglich, scheitert aber, wenn sie in gleicher

[120] Studienarbeiten werden trotzdem später in die Analyse von ehsStu einbezogen.

Form für mehrere Universitäten durchgeführt werden soll, an den unterschiedlichen Zusammensetzungen der Institutionen. Nur im Fall von DDC = 610 (Medizin) konnte an allen Einrichtungen, bei denen eine Medizinische Fakultät in den Daten auftritt, eine Zuordnung getroffen werden. Eine universitätsübergreifend zu verwendende fachliche Klassifikation kann anhand der vorhandenen Daten nicht stattfinden. Deshalb wurde dieses Vorhaben fallengelassen. Fakultäten werden für jedes IR individuell ausgewertet. Für die rein inhaltliche Klassifikation bietet sich die Verwendung der 10 Hauptklassen der DDC an. Da medizinische Publikationen, insbesondere Dissertationen, außer bei der Universität Stuttgart, aufgrund ihrer Anzahl eine besondere Rolle einnehmen, sollen sie als solche auch identifiziert werden können. Deshalb wird die Klasse „Technik, Medizin, angewandte Wissenschaften“ durch „Technik, angewandte Wissenschaften“ ersetzt und eine Klasse „Medizin“ eingeführt. Als gemeinsame inhaltliche Klassifikation wird die abgewandelte Einteilung der DDC in 11 Hauptklassen verwendet.

Tab. 16: Inhaltliche Klassifikation in 11 Klassen

0	Informatik, Informationswissenschaft, allgemeine Werke
1	Philosophie und Psychologie
2	Religion
3	Sozialwissenschaften
4	Sprache
5	Naturwissenschaften und Mathematik
6	Technik, angewandte Wissenschaften
61	Medizin
7	Künste und Unterhaltung
8	Literatur
9	Geschichte und Geografie

3.4 Aufbereitung der Daten

Metadaten und Nutzungsdaten lagen bei allen vier IR getrennt vor. Durch die Möglichkeit des direkten Zugriffs auf die Metadatenbank konnten bei edoc per SQL-Abfrage Metadaten und Nutzungsdaten bereits vor der Aufbereitung kombiniert werden. Insofern unterscheidet sich die Datenaufbereitung bei edoc von den anderen IR. Die Daten von ehsStu (Stuttgart), HeiDOK (Heidelberg) und SciDok (Saarbrücken) lagen in Form von Textfiles oder Microsoft-Excel-Files vor und waren in unterschiedlich viele Teilfiles aufgeteilt.

3.4.1 Aufbereitung der Metadaten

Da die zur Verfügung gestellten Metadaten verschieden aufbereitet waren, wurden sie zunächst für jedes IR extra in einer Tabelle gespeichert. Zu beachten waren dabei die unterschiedlichen Formate für das Datum, die in ein einheitliches Format überführt werden mussten.

OPUS-Metadaten (ehsStu, HeiDOK, SciDok)

Tab. 17: Spalten der OPUS-Metadaten-Tabellen vor der Zusammenführung mit edoc

Bezeichnung	Inhalt	Format
ID	Identifikation des Metadatensatzes	numerisch
Datum	Datum der Veröffentlichung in der Form Monat/Jahr	numerisch
Typ	OPUS-Publikationstyp*	numerisch
DDC	DDC 3-stellig	Zeichenkette
Institution	Fakultät oder Institut	Zeichenkette

* Typ ist abhängig von IR

Da die Werte für DDC ursprünglich nicht konsequent in 3-stelliger Form angegeben waren, mussten sie auf dieses Format transformiert werden. Daraus wurden 11 Inhaltsklassen gebildet. Der ursprüngliche Typ wurde in den gemeinsamen formalen Publikationstyp transformiert.

edoc

Bei edoc gab es einige Besonderheiten zu beachten. So enthält die Metadatenbank Datensätze, für die es zwar eine Frontdoor, aber keinen Volltext gibt. Diese Datensätze wurden ermittelt und gelöscht. Das ist wichtig, um sich bei der Anzahl der Publikationen in einer Gruppe auf Publikationen mit Volltexten zu beziehen. Des Weiteren können einer Publikation mehrere DDC-Werte zugeordnet werden, während bei den anderen IR der Wert eindeutig ist. Hier wurde so vorgegangen, dass zuerst für jeden DDC-Wert die Transformation in die 11 Klassen erfolgte. Dann wurde überprüft, ob sich die einzelnen DDC-Werte innerhalb einer Klasse befinden. Wenn ja, wurde die Klasse für die Publikation übernommen, wenn nicht, blieb der Wert leer. Bei den anderen IR wird der DDC-Wert in eine der 11 Klassen umgeschlüsselt[121].

Gemeinsames Metadatenformat

Nach der Aufbereitung der Metadaten für jedes IR wurden die Daten in einheitliche aufgebaute Tabellen überführt. Dabei wurden auch die originalen Bezeichnungen beibehalten, um jederzeit auf die Zuordnung zurückgreifen zu können, was die weitere Arbeit transparent macht. Die zweite Angabe einer Fakultät bei Dissertationen ist eine Besonderheit von HeiDOK. „Fakultaet“ bezeichnet das Ursprungsinstitut, welches nicht mit der zweiten Angabe übereinstimmen muss[122].

[121] Es wäre auch möglich gewesen, mehrere DDC-Werte zuzulassen und Publikationen mehr als einer der 11 Klassen zuzuordnen. Es handelt sich dabei aber nur um sehr wenige Publikationen. Aus diesem Grund und weil edoc damit einen Spezialfall darstellen würde, wurde darauf verzichtet.

[122] Als Ursprungsinstitut ist häufig eine zentrale Einrichtung angegeben.

Tab. 18: Metadaten aller vier IR

Name	Typ	Spalten-format	Dezimal-stellen	Inhalt
ID	Numerisch	4	0	ID des Metadatenssssatzes
Typ_ID_original	Numerisch	2	0	ID des originalen Publikationstyps
Typ_Label_original	String	50		Label des originalen Publikationstyps
DDC	String	3		Wert DDC der 100 Klassen der zweiten Ebene
Fakultaet	String	50		Fakultät oder andere Struktureinheit
Datum	Datum	10	0	Datum der Online-Publikation
P_Typ	String	1	0	Publikationstyp
I_Klasse	String	2	0	Inhaltsklasse
F_Diss	String	2		Fakultät bei Dissertationen

3.4.2 Aufbereitung der Nutzungsdaten

Die Daten wurden in unterschiedlicher Form geliefert. So gab es bei SciDok monatliche Nutzungsdaten für alle Teilfiles einzeln und nur für Publikationen, auf die im Erhebungszeitraum ein Zugriff erfolgte. Bei HeiDOK waren die Nutzungsdaten bereits summiert. Bei ehsStu enthielten die Nutzungsdaten den Wert 0, auch wenn die Publikation zum Erhebungszeitpunkt noch nicht vorhanden war. Außerdem waren die Nutzungsdaten bei ehsStu nicht monatlich erfasst, sondern halbjährlich bzw. für zwei Monate. Trotzdem sollte auf diese Daten nicht verzichtet werden und die für zwei Monate erhobenen Werte wurden verdreifacht. Durch die halbjährliche Erfassung kommt es im Vergleich zu den monatlichen Werten zu einer Nivellierung der Daten. Die Hochrechnung von 2 auf 6 Monate verstärkt die Effekte von Extremwerten und Ausreißern. Das Verfahren erwies sich aber als robust genug, um trotzdem nach Downloads signifikant unterschiedliche Gruppen bilden zu können[123]. Für edoc liegen zwar monatliche Nutzungsdaten für die Frontdoor und die Volltextformate PDF und HTML vor, wie bereits erwähnt, wird die Analyse auf die Downloads von PDF-Files beschränkt. Die Nutzungsdaten enthalten nur Werte, wenn Zugriffe erfolgten. Alle vorhandenen Nutzungsdaten wurden in Tabellen mit einheitlicher Struktur überführt. Enthalten sind teilweise fehlerhafte Downloadangaben, die korrigiert werden mussten. Das Datum der Erhebung des Downloads wurde auf den 1. Tag des Monats des Erhebungszeitraumes festgelegt. Für ehsStu, HeiDOK und SciDok ist Downloads als die Summe der Downloads der einzelnen PDF-Files im Erhebungszeitraum definiert.

[123] Downloads sind demzufolge bei ehsStu die Summe der Downloads in 6 Monaten. Der Erhebungszeitraum ist daher nicht einheitlich monatlich, wird aber durch den Monat des ersten Tages identifiziert.

Tab. 19: Nutzungsdaten aller vier IR

Name	Typ	Spalten-format	Dezimal-stellen	Inhalt
ID	Numerisch	4	0	ID des Metadatensatzes
D_access	Datum			1. Tag des Erhebungszeitraums
Downloads	Numerisch	2	0	PDF-Downloads im Erhebungszeitraum

3.4.3 Zusammenführung von Metadaten und Nutzungsdaten

Die Metadaten aller Publikationen wurden so vervielfältigt, dass für jeden vorhandenen Erhebungszeitraum ein Datensatz entstand, der durch ID und Monat oder Halbjahr eindeutig identifiziert wird. Über die Schlüssel ID und Datum wurden die Nutzungsdaten mit den Metadaten verbunden.

Um sicherzustellen, dass nur Downloads von Publikationen einbezogen werden, die während des gesamten Messzeitraums online waren, waren Korrekturen erforderlich. Bei der Entwicklung von NoRA wurden nur die Downloads berücksichtigt, für welche die Differenz von Erhebungszeitraum und Veröffentlichungsmonat größer oder gleich 1 ist[124]. Dadurch wurden auch die fehlerhaften Werte beseitigt, die durch Downloads = 0 von ehsStu auftraten. Andererseits mussten fehlende Werte durch 0 ersetzt werden, wenn kein Wert in den Nutzungsdaten vorhanden war, weil kein Zugriff im Erhebungszeitraum erfolgte. Das aus der Kombination von Metadaten und Nutzungsdaten entstandene File wird im Folgenden als Downloadfile bezeichnet. Da die Anzahl der Datensätze pro Publikation immer gleich der Anzahl der erfassten Erhebungszeiträume ist, kann die Auswahl der Gruppen, die analysiert werden sollen, immer vom prozentualen Anteil im Downloadfile erfolgen. Die Anzahl der Publikationen, die schließlich zu analysieren sind, kann durch den Quotient aus Anzahl der Datensätze und Anzahl der Erhebungszeiträume ermittelt werden. Das erleichtert die Analyse der Downloadfiles erheblich, da nur ein File verwendet werden muss.

[124] Das ist möglich, da Datum der Online-Publikation und Angabe des Messzeitraums numerisch sind und auf den 1. Tag des Erhebungsszeitraums gesetzt wurden.

3.5 Prinzipien und Komponenten von NoRA

Nach der Beschreibung des vorliegenden Datenmaterials und dessen Aufbereitung im vorangegangenem Abschnitt wird im Folgenden dargelegt, welche Eigenschaften die Daten aufweisen und welche Konsequenzen daraus zu ziehen sind. Konsequenzen ergeben sich zuerst aufgrund der Eigenschaften der Downloads, welche die Wahl des statistischen Verfahrens und damit den Kern von NoRA bestimmen. Im Anschluss daran wird gezeigt, wie das statistische Verfahren angewendet werden kann, um zum gewünschten Ergebnis, nämlich der Identifizierung von Gruppen[125] von Publikationen eines IR, die sich in ihren Downloads unterscheiden, zu gelangen. Danach wird darauf eingegangen, welche Überlegungen bei der Auswahl der Daten, die in die Analyse einzubeziehen sind, eine Rolle spielen und das Ergebnis als Festlegungen für die Datenauswahl zusammengefasst. Die Ergebnisse der drei Punkte

- die zu verwendenden statistische Testverfahren,
- die Anwendung der Testverfahren zur Bestimmung von Gruppen und
- die Festlegungen zur Datenauswahl

bilden zusammen die Analysemethode NoRA für IR.

3.5.1 Eigenschaften der Downloads und die Wahl der statistischen Verfahren

Dieser Abschnitt beschäftigt sich mit den statistischen Eigenschaften der Downloads. Ziel ist es, geeignete statistische Verfahren zu finden, die einen Vergleich der Downloads von Gruppen von Publikationen ermöglichen. Mit anderen Worten ausgedrückt: Es werden objektive Tests benötigt, aufgrund deren Ergebnis beurteilt werden kann, ob sich die Downloads von Gruppen von Publikationen signifikant unterscheiden und ein Vergleich anhand von zu bestimmenden Kennwerten sinnvoll ist. Welche Tests dafür geeignet sind, hängt von den statistischen Eigenschaften des Messwertes Downloads ab, die durch die Häufigkeitsverteilung charakterisiert werden. Zur Untersuchung der Häufigkeitsverteilung der Downloads wurde eine explorative Datenanalyse von edoc durchgeführt.

Da die charakteristischen Merkmale bereits bei der Verteilung der Downloads in einem Monat sichtbar sind und eine grafische Darstellung anhand des vollständigen Datenmaterials, welches bei edoc einen Zeitraum

[125] Gruppen von Publikationen werden auf der Grundlage der Kategorien von Merkmalen gebildet. Dabei kann eine Gruppe aus den Publikationen einer einzigen Kategorie oder der Zusammenfassung von Kategorien bestehen. In der Mathematischen Statistik wird eine Gruppe als Stichprobe bezeichnet, auch wenn es sich um die Gesamtheit und nicht nur eine Auswahl daraus handelt. Die Variable, die die Merkmale enthält, wird als Gruppenvariable bezeichnet. Werden mehrere Kategorien zu einer Gruppe zusammengefasst, wird in der Terminologie der Mathematischen Statistik eine neue Gruppenvariable gebildet.

von mehreren Jahren umfasst, nicht anschaulich ist, werden die Eigenschaften am Beispiel der Downloads von allen Publikationen eines bestimmten Monats erklärt. Wie zu erwarten war, zeigten bereits die ersten Auswertungen deutlich, dass es sich bei den Downloads um Daten handelt, die starken Schwankungen unterliegen. Die folgenden Abbildungen demonstrieren am Beispiel Januar 2007[126], wie monatliche Downloads typischerweise verteilt sind.

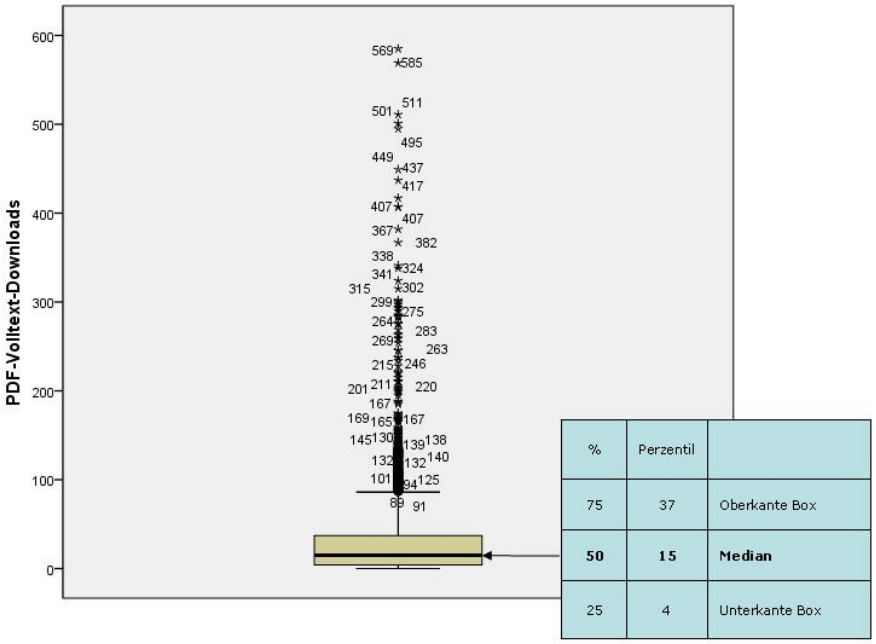


Abb. 9: Boxplot der PDF-Downloads, edoc 1/ 2007

Die „Box“ wird anhand von speziellen Kennwerten der Verteilung, den Perzentilen[127], gebildet. Die verwendeten Perzentile sind hierbei die Werte, für die gilt: bis zu 25 % bzw. 50 % bzw. 75 % aller Downloads liegen unterhalb dieses Wertes. Das 50 %-Perzentil ist der Median, welcher auch als zentrale Tendenz der Verteilung bezeichnet wird. Ausreißer und Extremwerte[128] sind die in der Grafik mit Werten versehenen Downloads. An der Grafik ist zu sehen, dass es viele Downloads gibt, die wahrscheinlich auf Robots zurückzuführen sind. In Ausnahmefällen können hohe Werte aber auch durch menschliche Zugriffe entstehen. Die Dichte und Anzahl der Ausreißer und Extremwerte muss immer zusammen mit der Anzahl der gültigen Werte betrachtet werden, die für diesen Monat 3909 beträgt. In der folgenden Tabelle sind weitere Perzentile angegeben. Daraus kann unter anderem entnommen werden, dass 99 % der Downloads unter dem Wert 238 liegen.

[126] Es wurde nur ein Monat ausgewählt, um die Grafiken übersichtlich gestalten zu können.

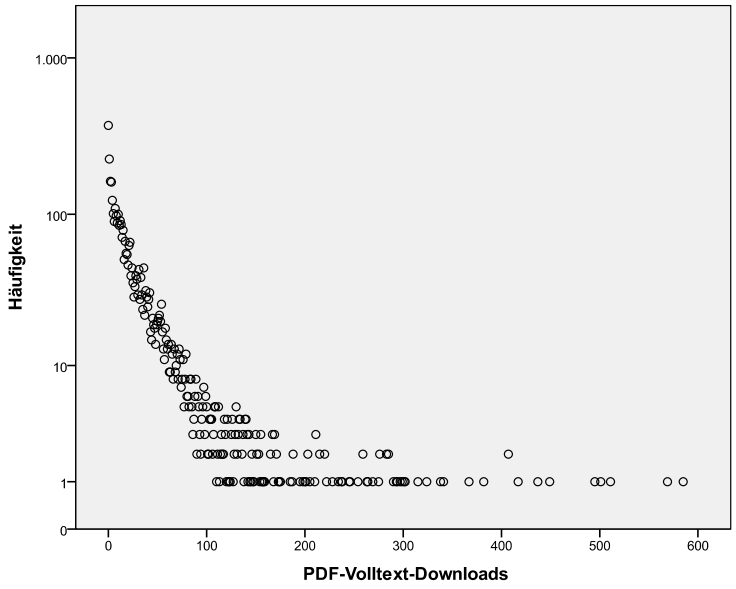
[127] Durch Perzentile werden die Daten in 100 Teile zerlegt. Das X-%-Perzentil ist der Wert, unterhalb dem X % aller gemessenen Werte liegen.

[128] Ausreißer sind Werte, die zwischen 1.5 und 3 Boxenlängen von der Oberkante aus liegen, Extremwerte sind die Werte, die über 3 Boxenlängen von der Oberkante aus liegen.

Tab. 20: Perzentile der PDF-Downloads, edoc 1/2007

Perzentile						
10 %	25 %	50 %	75 %	90 %	95 %	99 %
1	4	15	37	72	106	238

Die folgende Abbildung stellt die Verteilung der Downloads anhand der Häufigkeit der auftretenden Werte dar.



Die Y-Achse ist logarithmisch skaliert.

Abb. 10: Häufigkeiten der PDF-Downloads, edoc 1/2007

Es handelt sich um eine extrem schiefe Verteilung mit vielen Ausreißern und Extremwerten. Den Lageparameter, also einen Kennwert, der sich zur Unterscheidung eignet, einer solchen Verteilung durch das arithmetische Mittel zu schätzen, ist nicht sinnvoll. Um den Einfluss von Ausreißern und Extremwerten zu minimieren, wird zur Schätzung des Lageparameters der Median herangezogen[129].

[129] Der Lageparameter als charakteristisches Merkmal einer Häufigkeitsverteilung wird auch als zentrale Tendenz bezeichnet. Viele Vorgänge wie Messungen, deren Werte vom exakten Maß zufällig abweichen, werden durch eine Normalverteilung, auch als Gauß-Verteilung bezeichnet, gut modelliert. Sind Werte mehr oder weniger symmetrisch um einen Mittelwert verteilt, kann von einer Normalverteilung ausgegangen werden. In diesem Fall ist es sinnvoll, das arithmetische Mittel als Lageparameter zu schätzen und anzugeben. Bei einer Verteilung, bei der die Werte nicht symmetrisch um einen Mittelwert schwanken, ist dessen Verwendung nicht sinnvoll. Hier wird der Median als Lageparameter der Verteilung geschätzt und angegeben. Ein anschauliches Beispiel für die Verwendung des Medians ist die Berechnung der Armutsgefährdungsquote des Statistischen Bundesamtes. Als armutsgefährdet gilt, wer weniger als 60% des Medians des

In der folgenden Grafik werden die Häufigkeitsverteilungen der Downloads in zwei Kategorien eines Merkmals, hier der Inhaltsklasse, dargestellt und damit die Bedeutung der Lageparameter veranschaulicht. Als Inhaltsklassifikation wurde im Beispiel wegen der besseren Darstellbarkeit die Einteilung in die 10 DDC-Hauptklassen gewählt.

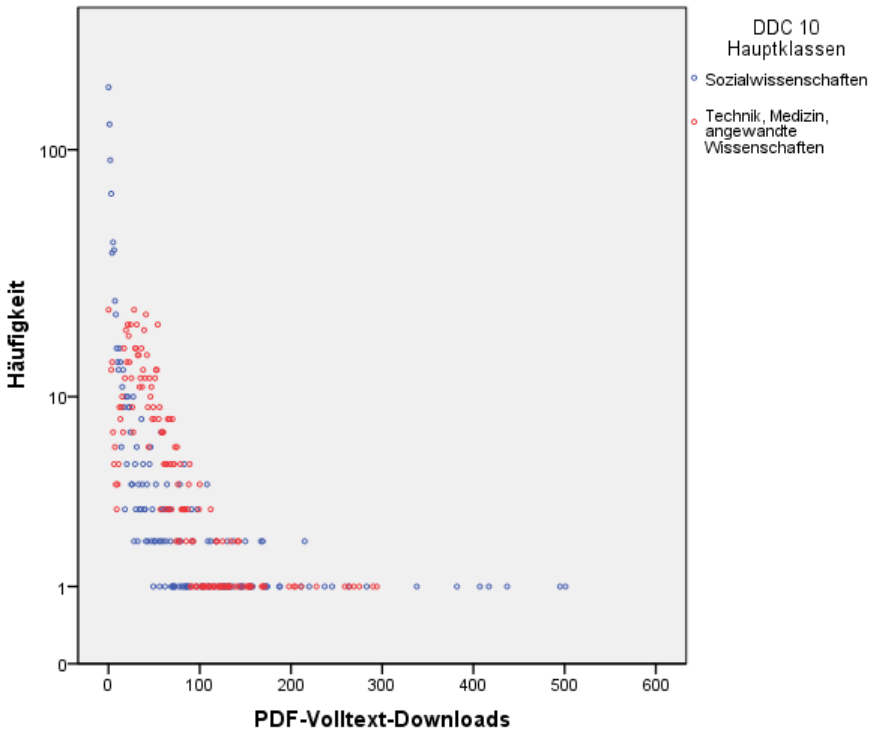
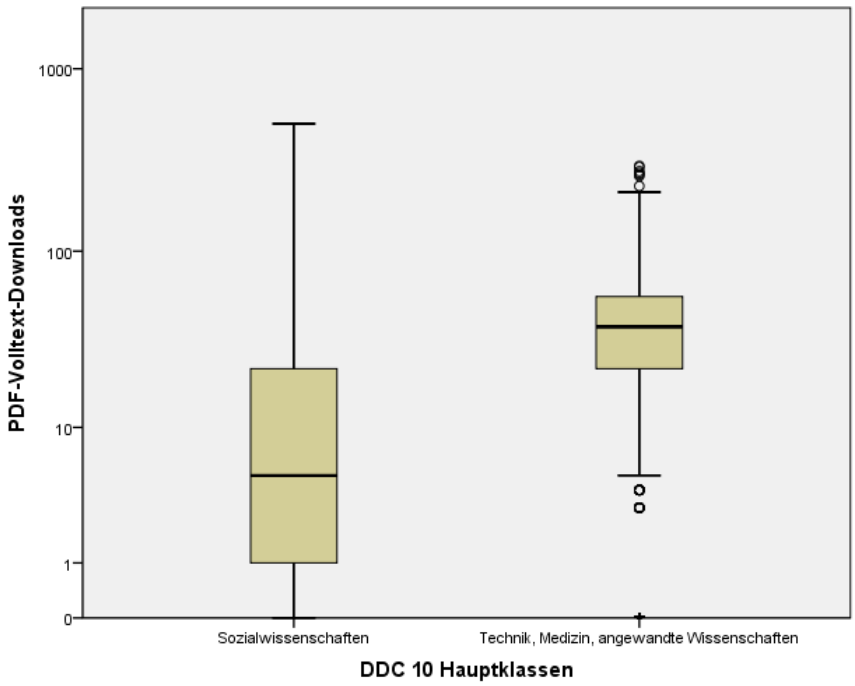


Abb. 11: Häufigkeiten der PDF-Downloads in zwei Kategorien von Publikationen, edoc 1/2007

Die Häufigkeiten der Downloads in der Klasse „Technik, Medizin, angewandte Wissenschaften“ im Vergleich zu „Sozialwissenschaften“ sind in Richtung höherer Werte verschoben. Es ist zu erwarten, dass sich die Lageparameter unterscheiden.

In der folgenden Boxplot-Darstellung können die Mediane direkt verglichen werden. Deutlich ist zu sehen, dass die Mediane stark voneinander abweichen.

Pro-Kopf-Einkommens der Bevölkerung zur Verfügung hat (Quelle http://www.amtliche-sozialberichterstattung.de/A1armutsgefaehrdungsquoten.html, gelesen 1.4.2011).



Die Y-Achse ist logarithmisch skaliert.

Abb. 12: Boxplots der PDF-Downloads in zwei Gruppen von Publikationen, edoc 1/2007

Damit ist aber noch nicht bekannt, ob sich die Verteilungen signifikant unterscheiden. Aufgrund der Abbildungen und dem Vergleich von Perzentilen kann das nur vermutet werden.

Tab. 21: Perzentile, edoc 1/2007

DDC 10 Hauptklassen	Perzentile						
	5 %	10 %	25 %	50 %	75 %	90 %	95 %
Sozialwissenschaften			1	5	22	66	111
Technik, Medizin, angewandte Wissenschaften	4	12	22	38	56	84	118

Zur Überprüfung muss ein Testverfahren der Mathematischen Statistik angewendet werden, mit dem entschieden wird, ob man tatsächlich von unterschiedlichen Verteilungen ausgehen kann. Welches Testverfahren verwendet wird, hängt davon ab, von welcher Annahme über die Häufigkeitsverteilung auszugehen ist. Bei Daten mit Häufigkeitsverteilungen mit der Schiefe, den Ausreißern und Extremwerten wie bei den Downloads kann nicht von einer Normalverteilung ausgegangen werden. In diesem Fall sind sogenannte nichtparametrische Verfahren auszuwählen. Dabei werden nicht die Werte selbst, sondern die Rangplätze, die die Werte innerhalb der zu analysierenden Grundgesamtheit oder Stichprobe einnehmen, analysiert. Der Rangplatz ist die Position, die der Wert einnimmt, wenn alle vorhandenen Werte der Größe nach sortiert werden. Da in der Regel Werte mehrfach auftreten, gibt es verschiedene Verfahren zur Ermittlung der Rangplätze.

Beurteilt werden soll, ob sich die Downloads, genauer gesagt die Verteilung der Downloads, in zwei oder mehr Gruppen von Publikationen unterscheiden. Als nichtparametrische Verfahren stehen der Mann-

Whitney-U-Test für zwei und der Test von Kruskal-Wallis für k unabhängige Gruppen oder Stichproben zur Verfügung[130]. Im Sinne der Mathematischen Statistik ist „Downloads“ die numerische intervallskalierte Testvariable und die Merkmale der Publikationen „Formaler Publikationstyp“, „Inhaltsklasse“ und „Fakultät“ sind kategoriale nominalskalierte Gruppenvariablen. Die Ausprägungen der Gruppenvariablen werden als Kategorien bezeichnet. Der Mann-Whitney-U-Test und der Test von Kruskal-Wallis sind Rangsummentests, d. h. die Summen der Ränge der Testvariablen werden ausgewertet. Dabei wird die Hypothese H_0: „Die Verteilung der Testvariablen ist in den Kategorien der Gruppenvariablen gleich“ geprüft. Wird die Standardeinstellung von SPSS für die Irrtumswahrscheinlichkeit p = 0,05 beibehalten und die Hypothese aufgrund der Teststatistik abgelehnt, kann mit einer Wahrscheinlichkeit von 95% davon ausgegangen werden, dass die Verteilungen der Testvariablen in den Kategorien der Gruppenvariablen nicht gleich sind. Eine Aussage mit einer Irrtumswahrscheinlichkeit von p <= 0,05 wird als signifikant bezeichnet. Wird die Hypothese H_0 abgelehnt, spricht man von signifikant unterschiedlichen Verteilungen oder gebraucht wie im Beispiel Downloads die verkürzte Formulierung „Die Downloads unterscheiden sich signifikant.“. Im Fall des Vergleichs der Verteilungen der Downloads in zwei Kategorien ergibt der Mann-Whitney-U-Test, dass die Hypothese H_0 abgelehnt werden kann und eine Darstellung in der folgenden Form sinnvoll ist.

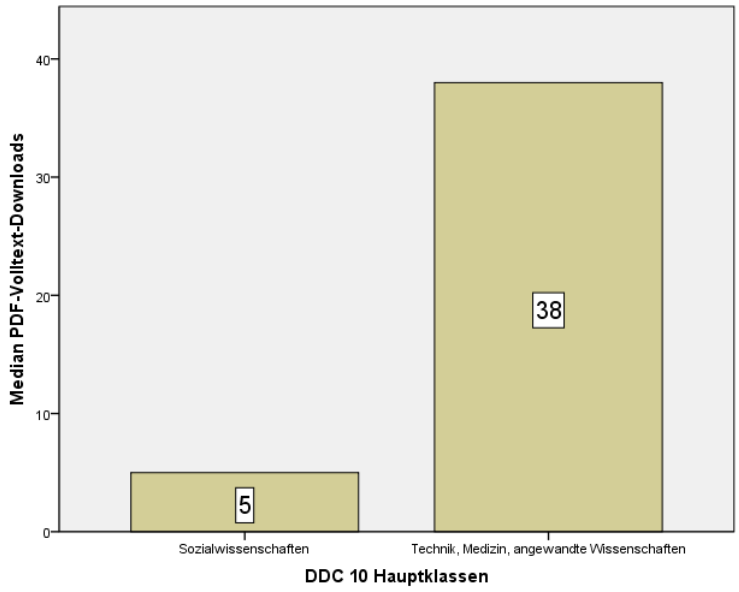


Abb. 13: Darstellung der Mediane der PDF-Downloads in zwei Kategorien von Publikationen, edoc 1/2007

Die explorative Datenanalyse führt damit zu dem Ergebnis, dass der Mann-Whitney-U-Test und der Kruskal-Wallis-Test anzuwenden sind, um zu beurteilen, ob sich die Verteilungen von Downloads in Gruppen von Publikationen signifikant unterscheiden. Wenn signifikante Unterschiede vorliegen, kann der Median als

[130] Der zuerst genannte Test ist nach Henry Mann und Donald Whitney benannt und wird auch als H-Test bezeichnet. Der Test von Kruskal-Wallis ist eine Erweiterung des Mann-Whitney-U-Test auf mehr als zwei Stichproben. Die Stichprobe sind die Downloads der Publikationen einer Gruppe von Publikationen.

Lageparameter der Verteilung von Downloads zum Vergleich von Gruppen von Publikationen herangezogen werden. Wenn das Testergebnis aussagt, dass die Hypothese beibehalten wird, unterscheiden sich die Downloads nicht signifikant und ein Vergleich ist nicht sinnvoll.

3.5.2 Praktische Durchführung der Signifikanztests

Nachdem festgestellt worden ist, welche Tests zur Prüfung der Hypothese H_0: „Die Verteilung der Testvariablen ist in den Kategorien der Gruppenvariablen gleich.“ geeignet sind, wird erläutert, wie diese anzuwenden sind, um zum gewünschten Ergebnis zu kommen. Ziel der Analyse ist es, Gruppen von Publikationen zu identifizieren, deren Downloads sich signifikant unterscheiden. Dabei steht nicht der Kontrast einzelner Kategorien, sondern die Gesamtheit der in Kategorien eingeteilten Publikationen im Vordergrund. Anders ausgedrückt: Es sind disjunkte Mengen von Kategorien so zu bilden, dass deren Downloads sich signifikant unterscheiden und Vergleiche der Mengen sinnvoll sind.

Werden nur die Downloads von zwei Gruppen bzw. Kategorien eines Merkmals verglichen, kann mit dem Mann-Whitney-U-Test beurteilt werden, ob es einen signifikanten Unterschied gibt. In der Regel sollen jedoch mehr als zwei Gruppen miteinander verglichen werden, weshalb der Test von Kruskal-Wallis anzuwenden ist. Wird durch den Test von Kruskal-Wallis die Hypothese bestätigt, sind für die Kategorien des getesteten Merkmals keine Unterschiede in den Downloads erkennbar und die Analyse muss für dieses Merkmal beendet werden. Führt der Test von Kruskal-Wallis zur Ablehnung der Hypothese H_0, bedeutet das, dass mindestens einer der paarweisen Vergleiche der Verteilungen in den Kategorien des getesteten Merkmals einen signifikanten Unterschied ergibt. Mehr wird zunächst nicht ausgesagt.

Um festzustellen, welche Kategorien sich signifikant unterscheiden, müssen alle paarweisen Vergleiche untersucht werden, deren Ergebnisse im Einzelnen angezeigt werden. Für welche Kategorien der Vergleich sinnvoll ist, hängt von der Anzahl der enthaltenen Fälle ab[131]. Deshalb muss zuerst überprüft werden, für welche Kategorien der Stichprobenumfang ausreichend groß ist. Nur diese Kategorien werden für den Test zugelassen. Im Idealfall erweisen sich alle Vergleiche als signifikant, was z. B.auf die Publikationstypen bei HeiDOK 2009 zutrifft. Getestet wurden die Downloads in den Kategorien „Dissertation“, „Abschlussarbeit“, Einzelpublikation“ und „Komplettpublikation“ der Gruppenvariablen „Formaler Publikationstyp“.

[131] Es besteht die Konvention, dass paarweise Vergleiche von zwei Stichproben vom Umfang n_1 und n_2 dann durchführbar sind, wenn gilt: $n_1 + n_2 > 30$ und (n1 > 20 oder $n_2 > 20$) (Kähler 2010, S. 415). An diese Konvention soll sich bei NoRA gehalten werden, da die individuelle Berechnung kritischer Werte für die Signifikanztests von den konkreten Stichprobengrößen abhängt und das Verfahren wesentlich komplizierter gestalten würde.

Die folgenden Grafiken zeigen, wie ein solches Testergebnis im sogenannten Model Viewer von SPSS aussieht.

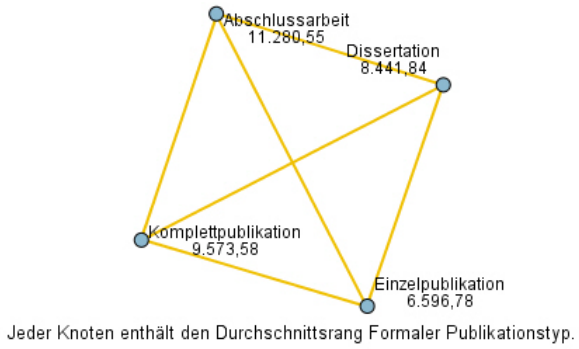


Abb. 14: Paarweise Vergleiche der Verteilung von Downloads nach Publikationstyp, Durchschnittsränge, HeiDOK 2009

Stichprobe1-Stichprobe2	Teststatistik	Standardfehler	Standard Teststatistik	Sig.	Anpassungssig.
Einzelpublikation-Dissertation	1.845,059	97,299	18,963	,000	,000
Einzelpublikation-Komplettpublikation	-2.976,804	200,590	-14,840	,000	,000
Einzelpublikation-Abschlussarbeit	4.683,772	198,638	23,579	,000	,000
Dissertation-Komplettpublikation	-1.131,746	185,734	-6,093	,000	,000
Dissertation-Abschlussarbeit	-2.838,713	183,625	-15,459	,000	,000
Komplettpublikation-Abschlussarbeit	1.706,968	253,943	6,722	,000	,000

Abb. 15: Paarweise Vergleiche der Verteilung von Downloads nach Publikationstyp, HeiDOK 2009

Die gelben Linien und Markierungen zeigen an, dass sich die Downloads signifikant unterscheiden.

In der Regel fallen nicht alle paarweisen Vergleiche signifikant aus. Je größer die Anzahl der Kategorien in den Gruppenvariablen ist, umso unübersichtlicher sind die Beziehungen der Kategorien. Ist k die Anzahl der Kategorien, so müssen $(k^2 - k)/2$ Paare überprüft werden. Bei 11 verschiedenen Kategorien sind das maximal 55 Vergleiche. Die Weiterarbeit hängt von der konkreten Fragestellung ab. Man kann eine bestimmte Kategorie auswählen und überprüfen, zu welchen anderen Kategorien signifikante Unterschiede bestehen, was aus dem Ergebnis in Form von Abbildung 13 leicht zu entnehmen ist. Ziel von NoRA ist jedoch, Aussagen über die Verteilung der Downloads in allen zugelassenen Kategorien zu machen, was anhand der grafischen Darstellung der Beziehungen als Netz bei größerer Anzahl von Kategorien nicht möglich ist. Ein Beispiel dafür ist der Test der Verteilungen der Downloads in den Kategorien der Inhaltsklasse. Der folgenden Tabelle kann entnommen werden, welche Kategorien für den Test aufgrund der Fallzahl zugelassen sind.

Tab. 22: Anzahl gültiger Fälle, edoc 1/2007

Inhaltsklasse	Anzahl gültige Fälle
0 Informatik, Informationswissenschaft, allgemeine Werke	179
1 Philosophie und Psychologie	48
2 Religion	8
3 Sozialwissenschaften	1042
4 Sprache	16
5 Naturwissenschaften und Mathematik	863
6 Technik, angewandte Wissenschaften	116
61 Medizin	794
7 Künste und Unterhaltung	14
8 Literatur	39
9 Geschichte und Geografie	33

Die Kategorien 2, 4 und 7 werden ausgeschlossen. In der grafischen Darstellung des Testergebnisses ist durch die blauen Linien zwar gut erkennbar, dass sich nicht alle Kategorien signifikant unterscheiden und diese zwei disjunkte Mengen bilden, die Zuordnung der Beschriftungen zu den Knoten ist aber nicht eindeutig und eine der Beschriftung fehlt völlig. Die Ausgabe des Testergebnisses vermittelt zwar einen ersten Eindruck, ist aber bezüglich der Mengenbildung von Kategorien mangelhaft.

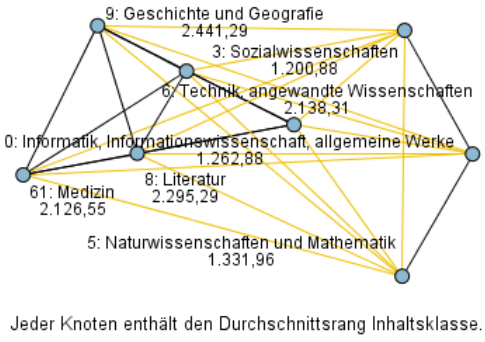


Abb. 16: Paarweise Vergleiche der Verteilung von Downloads nach Inhaltsklasse, Durchschnittsränge, edoc 1/2007

Um zu einem Ergebnis zu kommen, müssen alle paarweisen Vergleiche analysiert werden, wofür die folgende Ausgabe des Testergebnisses geeignet ist.

Stichprobe1-Stichprobe2	Teststatistik	Standardfehler	Standard Teststatistik	Sig.	Anpassungssig.
3: Sozialwissenschaften-0: Informatik, Informationswissenschaft, allgemeine Werke	62,003	72,669	,853	,394	1,000
3: Sozialwissenschaften-5: Naturwissenschaften und Mathematik	-131,083	41,339	-3,171	,002	,043
3: Sozialwissenschaften-61: Medizin	-925,668	42,310	-21,878	,000	,000
3: Sozialwissenschaften-6: Technik, angewandte Wissenschaften	-937,426	87,911	-10,663	,000	,000
3: Sozialwissenschaften-8: Literatur	-1.094,415	146,487	-7,471	,000	,000
3: Sozialwissenschaften-1: Philosophie und Psychologie	1.228,193	132,590	9,263	,000	,000
3: Sozialwissenschaften-9: Geschichte und Geografie	-1.240,408	158,805	-7,811	,000	,000
0: Informatik, Informationswissenschaft, allgemeine Werke-5: Naturwissenschaften und Mathematik	-69,081	73,766	-,936	,349	1,000
0: Informatik, Informationswissenschaft, allgemeine Werke-61: Medizin	-863,666	74,314	-11,622	,000	,000
0: Informatik, Informationswissenschaft, allgemeine Werke-6: Technik, angewandte Wissenschaften	-875,423	107,055	-8,177	,000	,000
0: Informatik, Informationswissenschaft, allgemeine Werke-8: Literatur	-1.032,412	158,716	-6,505	,000	,000
0: Informatik, Informationswissenschaft, allgemeine Werke-1: Philosophie und Psychologie	-1.166,190	145,988	-7,988	,000	,000
0: Informatik, Informationswissenschaft, allgemeine Werke-9: Geschichte und Geografie	-1.178,405	170,152	-6,926	,000	,000
5: Naturwissenschaften und Mathematik-61: Medizin	-794,585	44,167	-17,990	,000	,000
5: Naturwissenschaften und Mathematik-6: Technik, angewandte Wissenschaften	-806,343	88,820	-9,078	,000	,000
5: Naturwissenschaften und Mathematik-8: Literatur	-963,331	147,034	-6,552	,000	,000
5: Naturwissenschaften und Mathematik-1: Philosophie und Psychologie	1.097,109	133,194	8,237	,000	,000
5: Naturwissenschaften und Mathematik-9: Geschichte und Geografie	-1.109,324	159,310	-6,963	,000	,000
61: Medizin-6: Technik, angewandte Wissenschaften	11,758	89,276	,132	,895	1,000
61: Medizin-8: Literatur	-168,746	147,310	-1,146	,252	1,000
61: Medizin-1: Philosophie und Psychologie	302,524	133,499	2,266	,023	,656
61: Medizin-9: Geschichte und Geografie	-314,739	159,565	-1,972	,049	1,000
6: Technik, angewandte Wissenschaften-8: Literatur	-156,989	166,248	-,944	,345	1,000
6: Technik, angewandte Wissenschaften-1: Philosophie und Psychologie	290,767	154,143	1,886	,059	1,000
6: Technik, angewandte Wissenschaften-9: Geschichte und Geografie	-302,982	177,198	-1,710	,087	1,000
8: Literatur-1: Philosophie und Psychologie	133,778	193,624	,691	,490	1,000
8: Literatur-9: Geschichte und Geografie	-145,993	212,436	-,687	,492	1,000
1: Philosophie und Psychologie-9: Geschichte und Geografie	-12,215	203,103	-,060	,952	1,000

Abb. 17: Paarweise Vergleiche der Verteilung von Downloads nach Inhaltsklasse, edoc 1/2007

Danach sind die Vergleiche 3-0, 0-5, 61-6, 61-8, 61-1, 61-9, 6-8, 6-1, 6-9, 8-1, 8-9 und 1-9 nicht signifikant, wobei die dabei beteiligten Kategorien die beiden disjunkten Mengen M1 = (0, 3, 5) und M2 = (1, 6, 61, 8, 9) bilden. Für die Mengen M1 und M2 gilt, dass sich die Verteilungen der Downloads aller Kategorien aus M1 signifikant von den Downloads der Kategorien aus M2 unterscheiden. Zusammengefasst unterscheiden sich die Downloads von M1 signifikant von denen von M2. Das wird durch das folgende Ergebnis des Mann-Whitney-U-Tests demonstriert, welcher angewendet wird, da nur noch zwei Kategorien M1 und M2 analysiert werden.

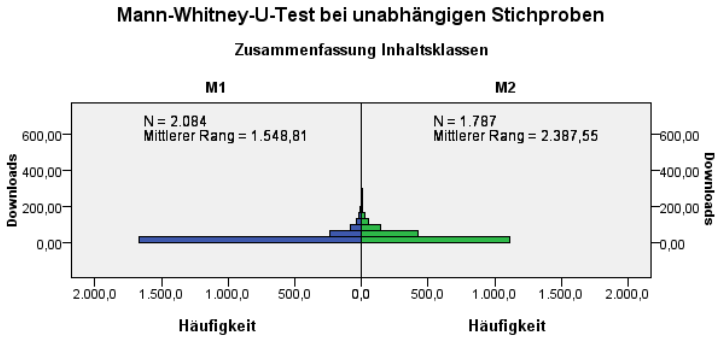


Abb. 18: Ergebnis des Mann-Whitney-U-Tests, edoc 1/2007

Die Reduzierung auf lediglich zwei Kategorien ist ein Spezialfall, wie sich im weiteren Verlauf zeigen wird. Es ist jedoch theoretisch möglich, dass nur eine Menge gebildet werden kann und dadurch, obwohl die Hypothese H_0 abgelehnt wurde, keine sinnvollen Vergleiche über alle Publikationen, sondern nur von einzelnen Kategorien möglich sind. Im vorliegenden Fall stellt die folgende Abbildung das Ergebnis der Analyse der Downloads für Januar 2007 dar.

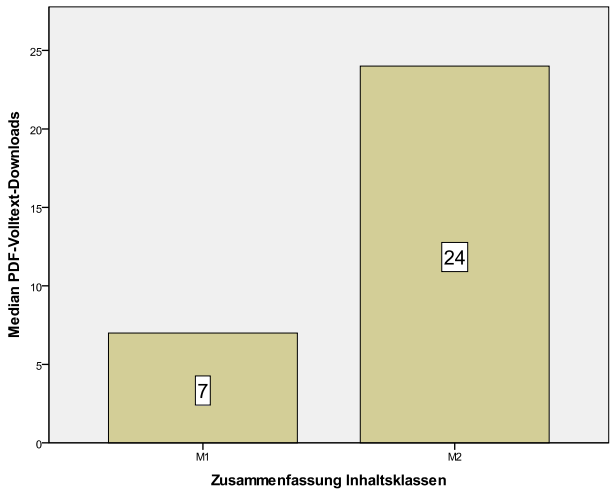


Abb. 19: Ergebnis der Analyse der Downloads nach Inhaltsklasse, edoc 1/2007

Auf die gleiche Art können alle paarweisen signifikanten Vergleiche der zugelassenen Kategorien von Inhaltsklasse dargestellt werden, ebenso die Vergleiche von M1 mit Kategorien von M2 und M2 mit den Kate-

gorien von M1. Die folgende Abbildung mit dem Vergleich der Downloads in den Naturwissenschaften mit den in der Menge M2 zusammengefassten Kategorien ist dafür ein Beispiel.

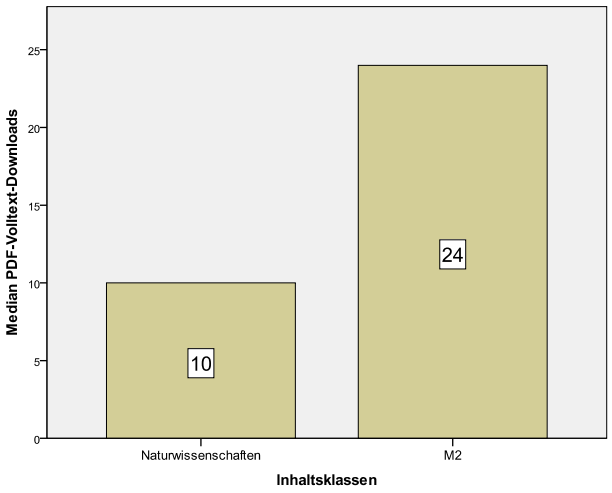


Abb. 20: Vergleich Downloads einer Kategorie mit der Zusammenfassung von Kategorien, edoc 1/2007

Die Verteilung der Downloads wird allerdings nicht nur durch den Median charakterisiert. Da die Signifikanz des Tests durch Rangsummen bestimmt wird, können paarweise Vergleiche signifikant und trotzdem die Mediane der beiden Gruppen gleich sein. Ein Beispiel dafür liefert der Test der Inhaltsklassen bei SciDok.

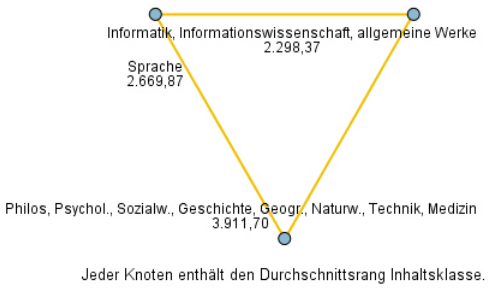


Abb. 21: Paarweise Vergleiche der Verteilung von Downloads nach Inhaltsklasse, Durchschnittsränge, SciDok 2009

Im Beipiel sind die Mediane von „Sprache“ und „Informatik, Informationswissenschaft, allgemeine Werke“ trotz signifikanter Unterschiede der Verteilungen gleich, die Durchschnittsränge unterscheiden sich jedoch. In solchen Fällen kann zwar der Vergleich der Verteilung der Downloads anhand der Durchschnittsränge der Gruppen erfolgen, eine Unterscheidung in zwei Gruppen hat aber keine praktische Bedeutung.

Durch die Anwendung des Mann-Whitney-U-Tests, des Tests von Kruskal-Wallis und der Auswertung der Testergebnisse ist es möglich, wenn signifikante Unterschiede der Häufigkeitsverteilungen der Downloads in Kategorien von Merkmalen existieren, diese zu ermitteln. Im Ergebnis können Gruppen von Kategorien und damit Gruppen von Publikationen gebildet werden, deren Downloads sich signifikant unterscheiden. Die Verteilungen der Downloads in diesen Gruppen können anhand des Lageparameters Median verglichen wer-

den. Wie bereits bemerkt, ist der Test von Kruskal-Wallis eine Erweiterung des Mann-Whitney-U-Tests auf mehr als zwei unabhängige Stichproben oder Gruppen. Da in der Regel mehr als zwei Gruppen zu vergleichen sind, kommt der Kruskal-Wallis-Test in der Mehrzahl der Fälle zur Anwendung und die Anwendung des Mann-Whitney-U-Tests bleibt die Ausnahme.

3.5.3 Die Auswahl der Daten

Nach der Festlegung auf geeignete Testverfahren und einen Anwendungsmodus zur Bestimmung von Gruppen von Publikationen mit signifikant unterschiedlichen Downloads muss noch geklärt werden, nach welchen Kriterien die Daten auszuwählen sind, die in die Analyse einbezogen werden, um im Ergebnis ein realistisches Bild der Nutzung der Publikationen zu erhalten. Im Folgenden wird nur auf monatliche Nutzungsdaten Bezug genommen, von denen bei NoRA typischerweise ausgegangen wird, und das Beispiel ehsStu mit Erhebungszeiträumen von 2 bzw. 6 Monaten nicht in die Erklärungen einbezogen. Dass NoRA trotzdem erfolgreich für die Daten aus Stuttgart angewendet und Ergebnisse erzielt werden, spricht für die Robustheit des Verfahrens.

Bei der Auswahl der Daten sind mehrere Aspekte zu berücksichtigen,

- die Größe der Gruppen von Publikationen,
- die Aktualität der Nutzungsdaten,
- die Aktualität der Metadaten und
- das Online-Alter der Publikationen,

die nicht isoliert betrachtet werden können. Wie sich zeigt, müssen Kompromisse gefunden werden, welche diese Aspekte weitgehend berücksichtigen und der Beschreibung der Nutzung des gesamten IR, dem Ziel von NoRA, am besten dienen.

Folgende Überlegungen spielen bei der Datenauswahl eine Rolle. Stünde nur die Aktualität im Vordergrund, wäre die Analyse auf die aktuellsten Nutzungsdaten zu beschränken und alle vorhandenen Publikationen einzubeziehen. Das würde dazu führen, dass es für viele Gruppen von Publikationen, vor allem, wenn es sich um IR mit relativ geringem Bestand handelt, eine zu niedrige Anzahl von Downloads gibt, um analysiert zu werden. Die minimale Anzahl von Werten wird einerseits durch die im vorangegangenen Abschnitt genannte Konvention für den Stichprobenumfang der verwendeten Tests mit $n > 20$ festgelegt. Wollte man lediglich die Daten des letzten zur Verfügung stehenden Monats analysieren, könnte man nur Gruppen mit mehr als 20 Publikationen zulassen, was eine unnötige Einschränkung wäre und außerdem den Verlust von Information bedeutet.

Aber auch wenn man die Anzahl der Monate auf Kosten der Aktualität erhöht, darf die Anzahl der Publikationen in einer Gruppe nicht uneingeschränkt verringert werden. Untersucht man Ausreißer und Extremwerte der Downloads genauer, stellt man fest, dass es zwei verschiedene Erscheinungsformen gibt. Bei der ersten Form treten extreme Werte zu verschiedenen Zeitpunkten und bei verschiedenen Publikationen auf, bei der zweiten Form gibt es für eine Publikation zu jedem Zeitpunkt einen Extremwert. Die erste Form rührt mit

hoher Wahrscheinlichkeit von einem Robot-Zugriff her. Die zweite Form kann als Ursache regelmäßige Robot-Zugriffe haben. Es kann sich aber auch um Publikationen handeln, bei denen keine Anzeichen für Robot-Zugriffe existieren, an denen aber generell ein großes Interesse besteht. Die Wahrscheinlichkeit, dass sich unter den Downloads einer Gruppe von Publikationen wegen ihrer geringen Größe Ausreißer und Extremwerte unverhältnismäßig stark häufen und das Ergebnis verfälschen, muss möglichst gering gehalten werden. In allen drei IR edoc, HeiDOK und SciDok existieren Publikationen mit regelmäßig weit erhöhten Downloads. Die Untersuchung von 6 aufeinanderfolgenden Monaten hatte zum Ergebnis, dass es Publikationen gibt, die in dieser Zeit immer Downloads aus dem Extrem- oder Ausreißerbereich aufweisen. Das sind bei edoc 1,2 %, bei SciDok 2,4 % und bei HeiDOK 2,6 % aller Publikationen. Daraus die Schlussfolgerung zu ziehen, dass für die Analyse zugelassene Gruppen von Publikationen nicht weniger als 3 % des Gesamtdatenbestand enthalten dürfen, ist nicht praktikabel, da dadurch wieder von vornherein viele Gruppen ausgeschlossen werden würden.

Die folgende Festlegung auf eine minimale Größe der Publikationsgruppe, die die Aspekte Aktualität der Nutzung und notwendige Stichprobengröße so versucht zu berücksichtigen, dass keine unverhältnismäßige Einschränkung der Anwendung von NoRA auf Teile des Bestandes an Publikationen entsteht, beruht auf den Gruppengrößen von edoc, HeiDOK und SciDok und damit auf Beständen von ca. 1000 bis 6000 Publikationen mit Volltexten. Danach werden Gruppen dann für die Analyse mit NoRA zugelassen, wenn ihr Anteil am zu untersuchenden Bestand, der eine Teilmenge der Publikationen des IR sein kann, wenn z. B. nur Dissertationen untersucht werden sollen, mindestens 1 % beträgt und mindestens 10 Publikationen enthalten sind. Zusammen mit der Verarbeitung der Downloads von 6 aufeinanderfolgenden Monaten konnten erwartungsgemäß nicht alle Gruppen, jedoch die für die Struktur des IR wesentlichen Gruppen in NoRA einbezogen werden.

Lässt man die Aktualität außer acht und analysiert größere Zeiträume mit NoRA, ist es unter Umständen nicht möglich, Gruppen von Publikationen zu bilden, deren Downloads sich signifikant unterscheiden. Hier kommt es offensichtlich zu Überlagerungen der Gruppeneffekte durch Einflüsse im Zeitverlauf und langfristige Änderungen im Bestand. Die genauere Untersuchung dieser sehr komplexen Zusammenhänge würde den Rahmen dieser Arbeit überschreiten. Eine Festlegung auf maximal 6 auszuwertende Monate führt bei edoc, ehsStu und Scidok dazu, dass die Gruppeneffekte durch zeitliche Einflüsse noch nicht überlagert werden.

Die Zugriffshäufigkeit ist unter anderem auch davon abhängig, wie lange eine Publikation bereits online auf dem IR verfügbar ist. Die folgende Grafik zeigt die signifikanten Unterschiede der Downloads, wenn man Publikationen, die maximal seit 6 Monaten online verfügbar sind, mit älteren vergleicht.

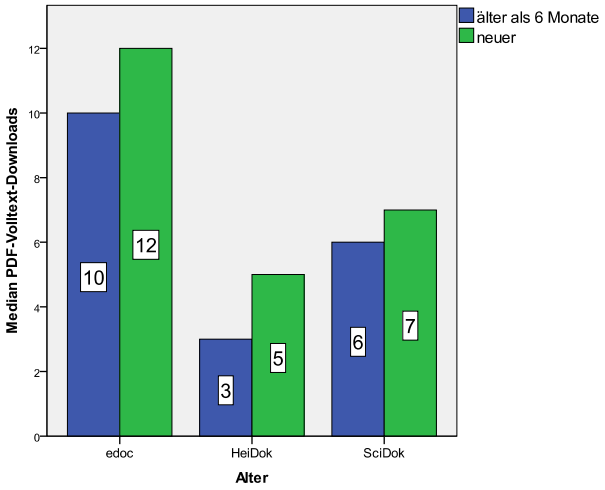


Abb. 22: Vergleich der Downloads nach Online-Alter der Publikationen

Ob ein anderes Online-Alter als 6 Monate zu größeren Unterschieden in den Downloads führt, wird im Rahmen dieser Arbeit nicht untersucht. Wichtig ist die Feststellung, dass es einen Einfluss gibt und dieser möglichst minimiert werden muss. Da die Veröffentlichung von Publikationen nicht kontinuierlich erfolgt, sondern sich die Anzahl in bestimmten Gruppen sprunghaft erhöht, kann der Effekt des Online-Alters den Gruppeneffekt überlagern und zu falschen Ergebnissen führen. Der Ausschluss neuer Publikationen kann das verhindern. Wie „neu“ für die Einbeziehung in NoRA sinnvoll zu definieren ist, kann wieder nur im Zusammenhang mit der gewünschten Aktualität und den erforderlichen Stichprobengrößen geklärt werden. Da es für drei IR diesen 6-Monats-Effekt gibt und ein Ausschluss von älteren Publikationen die Analyse zwar sicherer machen würde, die Aktualität des analysierten Bestandes und Umfang der Gruppen aber verringert, wird festgelegt, dass nur Downloads von Publikationen, die mindestens 7 Monate online sind, verwendet werden. Das könnte realisiert werden, indem man alle Downloads als fehlenden Wert behandelt, die entstanden sind, als das Online-Alter 6 Monate oder weniger betrug. Dann wäre jedoch die Prüfung, welche Gruppen groß genug für die Analyse sind, erheblich erschwert. Bezieht man grundsätzlich nur Publikationen ein, deren Online-Alter zum frühesten Erhebungsdatum bereits mindestens 7 Monate betrug, kann die Zulässigkeit der Gruppengröße einfach einer Häufigkeitstabelle entnommen werden. Mit dieser Festlegung schränkt man zwar die Aktualität des untersuchten Bestandes etwas ein, verhindert aber komplizierte Überprüfungsschritte, die dem Ziel einer möglichst einfach anwendbaren Analysemethode widerspricht.

Zusammengefasst werden folgende Festlegungen für die Auswahl der Daten, die in die Analyse mit NoRA einbezogen werden, getroffen:

- Gruppen von Publikationen sind für die Analyse zugelassen, wenn ihr Anteil am Datenbestand 1% nicht unterschreitet und die Anzahl an Publikationen mindestens 10 beträgt.
- Es werden Downloads von 6 aufeinanderfolgenden Monaten analysiert.
- Es werden Downloads von Publikationen analysiert, deren Online-Alter zum frühesten Erhebungsdatum mindestens 7 Monate beträgt

Mit diesen Festlegungen konnten für edoc, HeiDOK und SciDok anhand der Kategorien der Merkmale „Formaler Publikationstyp“, „Inhaltsklasse“ und „Fakultät“ Gruppen von Publikationen identifiziert werden, deren Downloads sich signifikant unterscheiden und die weitgehend die Struktur der IR abbilden.

4 Die Methode in 6 Schritten

Im 3. Kapitel ist dargestellt, wie die Daten der vier IR aufbereitet und anschließend zusammengeführt wurden. Auf der Grundlage dieses Datenmaterials erfolgte die Entwicklung der Grundprinzipien von NoRA. Im aktuellen Kapitel werden diese Prinzipien in einzelne Arbeitsschritte umgesetzt, die anhand eines Beispiels nachvollziehbar beschrieben werden.

Die Beschreibung der einzelnen Schritte und der Daten orientiert sich an der Terminologie des Programms SPSS, welches bereits für die Entwicklung der Methode verwendet wurde. Da es mit SPSS möglich ist, Daten über eine ODBC-Schnittstelle aus Datenbanken zu importieren, kann das gesamte erforderliche Datenmanagement in SPSS erfolgen. Voraussetzungen dafür sind die Installation eines ODBC-Treibers auf dem Computer, auf dem die Analyse durchgeführt werden soll und eine ODBC-fähige Datenbank des IR. SAS[132] ist ein weiteres Statistikprogramm, welches diese Möglichkeiten bietet, aber schwerer erlernbar und deshalb für diejenigen geeignet, die mit diesem Programm bereits vertraut sind. Für alle notwendigen Transformationen können auch andere geeignete Programme verwendet werden, z .B. Microsoft Excel oder mit Hilfe von SQL. Die grafischen Darstellungen können ebenso mit beliebigen anderen Programmen erzeugt werden. Ein Statistikprogramm ist lediglich zur Hypothesenprüfung mit dem Mann-Whitney-U-Tests und dem Tests von Kruskal-Wallis notwendig. Es bietet sich jedoch an, zumindest die Berechnung der Variablen der Klassifikationsmerkmale und die Analyse innerhalb eines einzigen Programms durchzuführen, welches dafür eine passende Umgebung bereitstellt. Befehle zur Berechnung von Variablen und Filtern und die Vergabe von Wertelabels für das Beispiel HeiDOK befinden sich in Anhang B in der Syntax von SPSS und können angepasst werden. Zur Erleichterung der Arbeit ist es sinnvoll, in den Häufigkeitstabellen sowohl Wert als auch Label der Variablen anzeigen zu lassen. Screenshots der Menüs zur Durchführung der Tests zur Hypothesenprüfung und der Erstellung der Grafiken in SPSS befinden sich ebenfalls im Anhang B.

Die 6 Schritte der Methode zur Analyse eines universitären IR werden am Beispiel des formalen Publikationstyps und der Fakultäten von Dissertationen der Daten von HeiDOK 2009 erläutert. Alle im Folgenden durchgeführten Auswertungen sind Beispiele für die vorhandenen Möglichkeiten und dienen als Anregung und Demonstration. Es müssen nicht zwangsläufig die gleichen Klassifikationen verwendet werden, als formale und inhaltliche Klassifikation bieten sich jedoch Publikationstyp und die vorgestellten Inhaltsklassen an. Deshalb wird die Aufbereitung der Daten nach diesen Merkmalen im ersten Schritt berücksichtigt.

Die Extraktion der Daten kann nicht in allgemeiner Form beschrieben werden, da Metadaten und Nutzungsdaten bei den IR in unterschiedlichen Formaten und auf unterschiedliche Tabellen verteilt gespeichert werden. Der Anwender der Methode kann sich jedoch an der Aufbereitung der Daten der Studienteilnehmer orientieren (siehe Abschnitt 3.4).

[132] SAS: Business Analytics Software, http://www.sas.com/offices/europe/germany/.

4.1 Schritt 1: Erstellung des Metadatenfiles

Das Ausgangsmaterial bilden Metadaten der Publikationen des IR, welche in unterschiedlicher Granularität vorliegen können. Ergebnis des ersten Schrittes ist ein Metadatenfiles, welches den aktuellen Bestand an Publikationen mit PDF-Volltexten, das Datum der Online-Publikation und formale und inhaltliche Merkmale in den Formen enthält, die der gewünschten Analyse entsprechen. Bereits hier sollte darauf geachtet werden, die Anzahl der Kategorien der Merkmale auf ein vernünftiges Maß zu begrenzen.

Wenn die inhaltliche Klassifikation auf dem Wert des DDC beruhen soll, muss dieser möglicherweise in eine 3-stellige Form umcodiert werden. Die Variablen P_Typ und I_Klasse werden aus dem originalen Publikationstyp und DDC gebildet. Die Möglichkeit der Vergabe von Labels für die Variablen sollte man nutzen, da es zur besseren Übersicht beiträgt und die Erstellung von Grafiken vereinfacht. Nach der Berechnung von P_Typ können alle Metadatensätze, die keinen Volltext repräsentieren, gelöscht oder gefiltert werden. Empfohlen wird die Löschung, da im Weiteren mehrfach Filter gebildet werden müssen und dadurch die Übersicht, was gerade gefiltert wird, erschwert werden würde.

Die folgende Tabelle beschreibt den Aufbau des Metadatenfiles zur Analyse der Publikationen nach zwei formalen und einem inhaltlichen Merkmal.

Tab. 23: Aufbau des Metadatenfiles

Name	Typ	Spalten-format	Dezimal-stellen	Inhalt
ID	Numerisch	4	0	Identifikation des Metadatensatzes
Datum	Datum	10	0	Datum der Online-Publikation
P_Typ	String	1		Formaler Publikationstyp
I_Klasse	String	2		Inhaltsklasse
F_Diss	String	2		Fakultät von Dissertationen

Die Länge der String-Variablen richtet sich nach der Anzahl der zu codierenden Kategorien. Das Datumsformat muss auf einer numerischen Codierung beruhen, damit ein Vergleich mit dem Datum des Erhebungsmonats möglich ist. P_Typ, I_Klasse und F_Diss sind kategoriale Gruppenvariablen.

4.2 Schritt 2: Analyse des Metadatenfiles

Analysiert wird das im ersten Schritt erstellte Metadatenfile. Ziel ist es, den Bestand an Publikationen und dessen Entwicklung grafisch darzustellen. Eine Gesamtübersicht der Entwicklung enthält die folgende Abbildung.

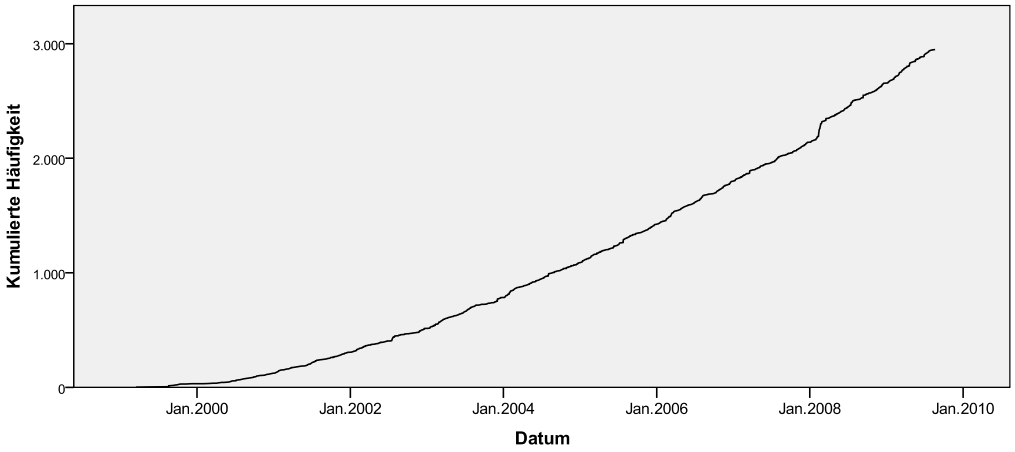


Abb. 23: Entwicklung des Bestandes an Publikationen, HeiDOK 7/2009

Einen ersten Überblick über die Verteilung des formalen Publikationstyps, Stand Juli 2009, ergeben Häufigkeitstabellen.

Tab. 24: Häufigkeiten Publikationstyp, HeiDOK 7/2009

P_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	2136	72,4	73,3	73,3
	E Einzelpublikation	529	17,9	18,2	91,5
	C Abschlussarbeit	125	4,2	4,3	95,8
	F Komplettpublikation	118	4,0	4,1	99,8
	B Habilitation	5	,2	,2	100,0
	Gesamt	2913	98,8	100,0	
Fehlend		36	1,2		
Gesamt		2949	100,0		

Aus den Häufigkeitstabellen kann abgelesen werden, welche Kategorien maximal analysiert werden können. Der Publikationstyp B wird in den folgenden Grafiken weggelassen, da er bereits im Metadatenfile mit 5 zu wenig Fälle aufweist.

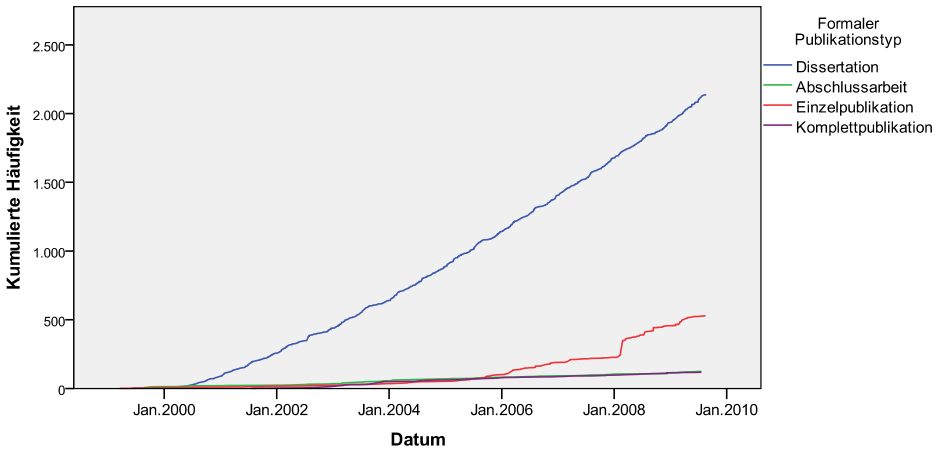


Abb. 24: Entwicklung des Bestandes an Publikationen nach Publikationstyp, HeiDOK 7/2009

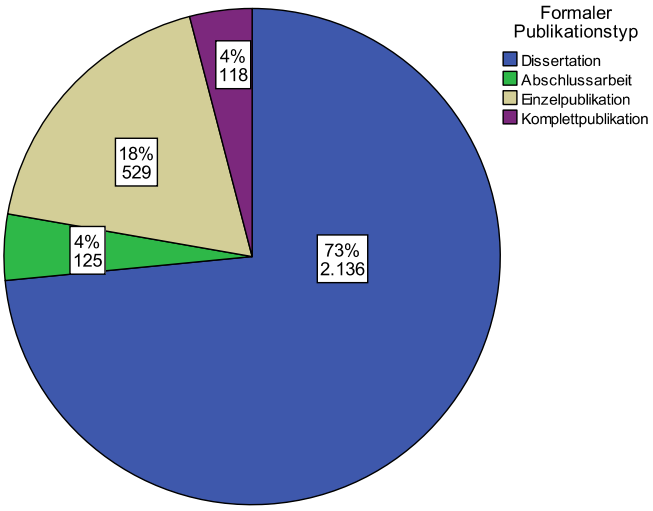


Abb. 25: Verteilung der Publikationen nach Publikationstyp, HeiDOK 7/2009

Aus den aufbereiteten Metadaten werden mit einfachsten Mitteln Grafiken erzeugt, die einen Überblick über die Entwicklung des Bestandes an Publikationen mit Volltextdokumenten geben. Hier ist ersichtlich, dass Dissertationen den überwiegenden Anteil an Publikationen darstellen und ihre Anzahl kontinuierlich wächst. Der Anteil an Einzelpublikationen wächst, aber auf einem viel geringerem Niveau.

Mit den gleichen Mitteln können weitere Übersichtstabellen und Grafiken unter Verwendung anderer Variablen erzeugt werden.

4.3 Schritt 3: Erstellung des Downloadfiles

Die ideale Voraussetzung sind monatlich aggregierte Downloads über den Zeitraum eines halben Jahres. Sollten solche Nutzungsdaten nicht vorhanden sein, können auch anders aggregierte und andere Zeiträume erfassende Downloads wie im Fall von ehsStu verwendet werden. Das hat Konsequenzen für die zur Analyse zugelassene Anzahl der Publikationen und muss im folgenden Schritt beachtet werden. Je länger die einzelnen Erhebungszeiträume und der gesamte Zeitraum sind, umso größer ist die Wahrscheinlichkeit, dass keine Publikationsgruppen mit signifikanten Unterschieden der Downloads ermittelt werden können.

Die Nutzungsdaten des IR müssen so aufbereitet sein, dass der Monat (oder ein anderer Erhebungszeitraum) ein Datumsformat aufweist, welches mit dem Datumsformat des Datums der Online-Publikation übereinstimmt. Die Variable D_access enthält das Datum des ersten Tages des Monats, in dem die Downloads gezählt wurden. Ist der Erhebungszeitraum nicht ein Monat, so erhält die Variable den Datumswert des ersten Tages des Erhebungszeitraumes. Der Aufbau des Nutzungsdatenfiles ist identisch mit der Beschreibung in Tabelle 14. Falls Nutzungsdaten für mehr als 6 Erhebungszeiträume vorhanden sind, werden die letzten 6 ausgewählt (Anzahl m = 6). Für jeden Datensatz des Metadatenfiles werden m Datensätze generiert und mit den Schlüsseln ID und D_access versehen. Metadaten und Nutzungsdaten werden über diese Schlüssel verbunden. Wegen der Anzahl Metadatensätze n = 2949 erhält man ein Downloadfile mit n·m = 17694 Datensätzen. Anschließend werden alle Publikationen gefiltert oder gelöscht, die später als 6 Monate vor dem ersten Erhebungszeitraum 2/2009 online veröffentlicht wurden, d. h. analysiert werden nur Publikationen, für die Datum < 8/2009 gilt. Normalerweise befinden sich im Nutzungsdatenfile nur Daten, für die es Zugriffe gegeben hat. In diesem Fall müssen nicht vorhandene Werte der Variablen Downloads auf 0 gesetzt werden. Das Downloadfile, welches der Ausgang für die nächsten Schritte ist, hat folgenden Aufbau:

Tab. 25: Aufbau des Downloadfiles

Name	Typ	Spalten-format	Dezimal-stellen	Inhalt
ID	Numerisch	4	0	Identifikation des Metadatensatzes
Datum	Datum	10	0	Datum der Online-Publikation
P_Typ	String	1		Formaler Publikationstyp
I_Klasse	String	2		Inhaltsklasse
F_Diss	String	2		Fakultät von Dissertationen
D_access	Datum			1. Tag des Erhebungszeitraums
Downloads	Numerisch	2	0	PDF-Downloads im Erhebungszeitraum

Die Variable „Downloads“ ist die Testvariable.

4.4 Schritt 4: Bestimmung der zugelassenen Kategorien

Anhand der aus dem Downloadfile erzeugten Häufigkeitstabellen wird ermittelt, für welche Kategorien der Merkmale von Publikationen die Downloads analysiert werden können. Ausschlaggebend hierfür ist die Anzahl der Fälle in der Kategorie, die 1 % nicht unterschreiten darf und mindestens 10 Fälle im Fall von 6 Downloadwerten pro Publikation enthalten muss. Hier wird die Analyse für den Publikationstyp und die Fakultät von Dissertationen demonstriert. Bei der Analyse nach Inhaltsklassen wird analog vorgegangen. Im Anhang ist jeweils die Berechnung des Filters für die Auswahl der zu analysierenden Kategorien in der Terminologie von SPSS angegeben.

Fomaler Publikationstyp

Tab. 26: Häufigkeiten Publikationstyp im Downloadfile, HeiDOK 2009

P_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	12816	72,4	73,3	73,3
	B Habilitation	30	,2	,2	73,5
	C Abschlussarbeit	750	4,2	4,3	77,8
	E Einzelpublikation	3174	17,9	18,2	95,9
	F Komplettpublikation	708	4,0	4,1	100,0
	Gesamt	17478	98,8	100,0	
Fehlend		216	1,2		
Gesamt		17694	100,0		

Aus der Häufigkeitstabelle geht hervor, dass es nur 5 Habilitationen gibt, da im Downloadfile pro Publikation soviel Datensätze wie Monate vorhanden sind. Der in der Untersuchung der Nutzungsdaten ermittelte Wert von 1 % (siehe Abschnitt 3.5.3) der gesamten Datensätze wird mit 0,2 % unterschritten. Damit ist diese Gruppe von vornherein zu klein, um deren Nutzungsdaten auszuwerten und wird daher ausgeschlossen. Alle anderen Publikationstypen können analysiert werden.

Fakultäten von Dissertationen

Möchte man z. B. die Nutzung der Dissertationen verschiedener Fakultäten vergleichen, sind die folgenden Häufigkeiten ausschlaggebend.

Tab. 27: Häufigkeiten Fakultät von Dissertationen im Downloadfile, HeiDOK 2009

F_Diss Fakultät Dissertation					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 Fak. für Biowiss.	3426	26,7	28,5	28,5
	10 Neuphilologische Fakultät	132	1,0	1,1	29,6

	11 Philosophische Fakultät	486	3,8	4,0	33,6
	12 Theologische Fakultät	96	,7	,8	34,4
	13 Zentrale und Sonstige Einrichtungen	42	,3	,3	34,7
	2 Fak. f. Chemie und Geowiss.	1950	15,2	16,2	50,9
	3 Fak. f. Mathematik und Informatik	630	4,9	5,2	56,2
	4 Fak. f. Physik und Astronomie	3924	30,6	32,6	88,8
	5 Fak. f. Verhaltens- und Empirische Kulturwiss.	756	5,9	6,3	95,1
	6 Fak. f. Wirtschafts- und Sozialwiss.n	492	3,8	4,1	99,2
	7 Juristische Fakultät	6	,0	,0	99,2
	8 Medizinische Fakultät Heidelberg	72	,6	,6	99,8
	9 Medizinische Fakultät Mannheim	24	,2	,2	100,0
	Gesamt	12036	93,9	100,0	
Fehlend		780	6,1		
Gesamt		12816	100,0		

Alle rot markierten Fakultäten werden wegen geringer Fallzahlen von der Analyse ausgeschlossen. Die übrigen sind für die Analyse mit NoRA zugelassen. An diesem Beispiel wird deutlich, wie sich die Anzahl von Kategorien auf die Stichprobengrößen auswirkt. Es kann von vornherein überlegt werden, wie die Anzahl von Kategorien durch Zusammenfassungen sinnvoll verringert werden kann. Das ist jedoch eine Ermessensfrage und kann zu Informationsverlust führen, wie später zu sehen ist.

4.5 Schritt 5: Signifikanztests für zugelassene Kategorien

Dieser Schritt stellt den Kern von NoRA dar, weil in ihm überprüft wird, ob es überhaupt signifikante Unterschiede zwischen den Downloads der ausgewählten Kategorien der Merkmale von Publikationen gibt. Falls signifikante Unterschiede vorliegen, werden aus den vorher zur Analyse ausgewählten Kategorien Zusammenfassungen zu Gruppen oder Mengen von Kategorien gebildet, die sich paarweise signifikant unterscheiden.

Geprüft wird die Hypothese H_0: Die Verteilungen der Downloads in den Gruppen unterscheiden sich nicht. Wird die Hypothese H_0 abgelehnt, kann davon ausgegangen werden, dass sich die Verteilungen unterscheiden und es sinnvoll ist, die Lageparameter Median zu vergleichen. Wird dagegen die Hypothese im Test bestätigt, erübrigt sich die Weiterarbeit. In der Regel kommt in Schritt 5 der Test von Kruskal-Wallis zur Anwendung, da mehr als 2 Kategorien überprüft werden sollen, im Ausnahmefall von 2 Kategorien ist der Mann-Whitney-U-Test auszuwählen. In SPSS kann es dem Programm überlassen werden, welcher der beiden Tests in Abhängigkeit von der Anzahl der Kategorien benutzt wird. Die Hypothese H_0 wird getestet, indem im Menü

„Analysieren“ der Punkt „Nichtparametrische Tests, unabhängige Stichproben“ gewählt wird. Alle Standardeinstellungen wie das Signifikanzniveau 0,05 können beibehalten werden und nur die Testvariable „Downloads“ und die Gruppenvariable sind auszuwählen.

Formaler Publikationstyp

Der Filter filter_P_Typ wird auf das Downloadfile angewendet, wodurch der Publikationstyp „Habilitation“ ausgeschlossen wird, und anschließend der Test für die Variable Downloads und die Variable P_Typ durchgeführt. Das Testergebnis wird in folgender Form angezeigt:

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von Downloads ist in den Kategorien von Formaler Publikationstyp identisch.	Kruskal-Wallis-Test bei unabhängigen Stichproben	,000	Nullhypothese ablehnen

Abb. 26: Ergebnis des Kruskal-Wallis-Tests für Publikationstyp in SPSS, HeiDOK 7/2009

Wird die Hypothese H_0 abgelehnt, gibt es mindestens bei einem der paarweisen Vergleiche einen signifikanten Unterschied der Downloads. Im speziellen Fall gibt bei allen paarweisen Vergleichen der zugelassenen Kategorien signifikante Unterschiede der Verteilungen. In SPSS wird dieses Ergebnis in Form der folgenden Abbildung dargestellt.

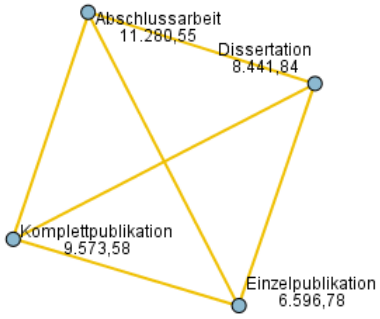


Abb. 27: Paarweise Vergleiche Publikationstyps, Durchschnittsränge, HeiDOK 7/2009

Gelbe Verbindungslinien zwischen den Kategorien bedeuten, dass sich die Verteilungen der Downloads signifikant unterscheiden. Die Zahlen geben den Durchschnittsrang in den Kategorien an.

Fakultäten von Dissertationen

Der Filter filter_P_F_Diss wird auf das Downloadfile angewendet, damit nur ausreichend große Publikationsgruppen in die Analyse eingehen, und anschließend der Test für die Variable Downloads und die Variable F_Diss durchgeführt.

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von Downloads ist in den Kategorien von Fakultät Dissertation identisch.	Kruskal-Wallis-Test bei unabhängigen Stichproben	,000	Nullhypothese ablehnen

Abb. 28: Ergebnis des Kruskal-Wallis-Tests, Fakultäten von Dissertationen, HeiDOK 7/2009

Hier gibt es bei 5 von 28 paarweisen Vergleichen der Kategorien keine Signifikanz. In diesem Fall müssen Kategorien zusammengefasst werden. Wie das zu tun ist, kann man den paarweisen Vergleichen entnehmen.

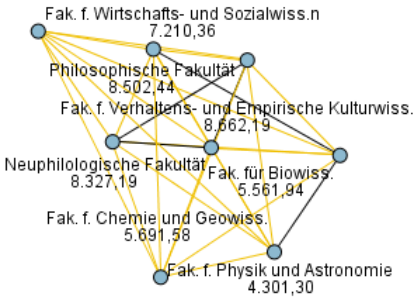


Abb. 29: Paarweise Vergleiche, Fakultät von Dissertationen vor der Zusammenfassung von Fakultäten, Durchschnittsränge, HeiDOK 7/2009

Eine blaue Verbindungslinie zeigt an, dass der paarweise Vergleich nicht signifikant ist. Im Anhang befindet sich die vollständige Liste. Aus der Analyse der paarweisen Vergleiche ergibt sich, welche Fakultäten zusammengefasst werden müssen, um bei allen paarweisen Vergleichen signifikante Ergebnisse zu erhalten.

Nicht signifikant sind die Vergleiche:

3 Fak. f. Mathematik und Informatik - 1 Fak. für Biowiss.

3 Fak. f. Mathematik und Informatik - 2 Fak. f. Chemie und Geowiss.

10 Neuphilologische Fakultät - 11 Philosophische Fakultät

10 Neuphilologische Fakultät - 5 Fak. f. Verhaltens- und Empirische Kulturwiss.

11 Philosophische Fakultät - 5 Fak. f. Verhaltens- und Empirische Kulturwiss.

Die an diesen Vergleichen beteiligten Fakultäten bilden disjunkte Mengen M1 = (1,2,3) und M2 = (5,10,11), wobei gilt, dass alle paarweisen Vergleiche zwischen den Elementen von M1 und M2 signifikant sind. Zur Überprüfung des Ergebnisses kann erneut der Test von Kruskal-Wallis ausgeführt werden, für den eine neue Gruppenvariable F_Diss_Sign für Fakultäten gebildet wird, bei der die in M1 und M2 enthaltenen Kategorien

jeweils zusammengefasst werden. Die Anweisung zur Zusammenfassung der Kategorien befindet sich im Anhang. Für die Kategorien von F_Diss_Sign fallen alle paarweisen Vergleiche signifikant aus.

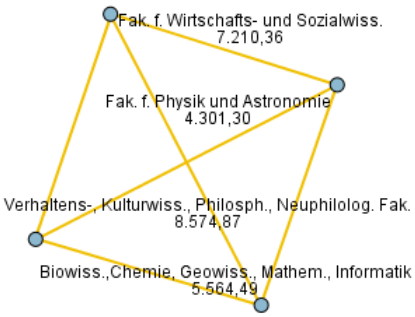


Abb. 30: Ergebnisse der paarweisen Vergleiche von Verteilungen von Downloads für Fakultäten und Fakultätsgruppen von Dissertationen, Durchschnittsränge, HeiDOK 2009

Eine andere Möglichkeit wäre auch die Zusammenfassung von Fakultäten von vornherein gewesen. Wahrscheinlich wären dabei spezifische Unterscheidungsmöglichkeiten für zwei Fakultäten, die nicht mit anderen zusammengefasst werden, nicht in Betracht gezogen worden und somit verlorengegangen, so dass das aufwendigere demonstrierte Verfahren Vorteile bietet.

4.6 Schritt 6: Grafische Darstellung der Ergebnisse

Im letzten Schritt werden die Ergebnisse von NoRA in Form von Grafiken dargestellt, die eine Gegenüberstellung der Unterschiede der Downloads und der Entwicklung des Bestandes in den gebildeten Publikationsgruppen ermöglicht. Selbst in ein- und derselben Programmumgebung gibt es viele Möglichkeiten, Grafiken zu erstellen. Sinnvoll ist es, die Gruppen in der Reihenfolge ihrer Mediane anzuordnen. Die hier vorgestellte Variante ist nur eine von vielen. Sie wurde gewählt, weil damit die unterschiedlichen Verteilungen in den Gruppen gut dargestellt werden können. Im vorliegenden Fall von SPSS wurden die Mediane über das Menü „Diagramme“ berechnet. Es ist ebenso möglich, Mediane in Microsoft Excel (ab Version 2007) zu berechnen und die Grafiken in diesem Programm zu erstellen.

Formaler Publikationstyp

Von den 5 Kategorien wurde „Habilitation“ aufgrund geringer Fallzahl aus der Analyse ausgeschlossen. Die paarweisen Vergleiche der Downloads der übrigen Kategorien fielen alle signifikant aus, so dass keine Kategorien zusammengefasst werden mussten. In diesem Fall können für alle übrigen 4 Kategorien die Mediane berechnet und grafisch dargestellt werden.

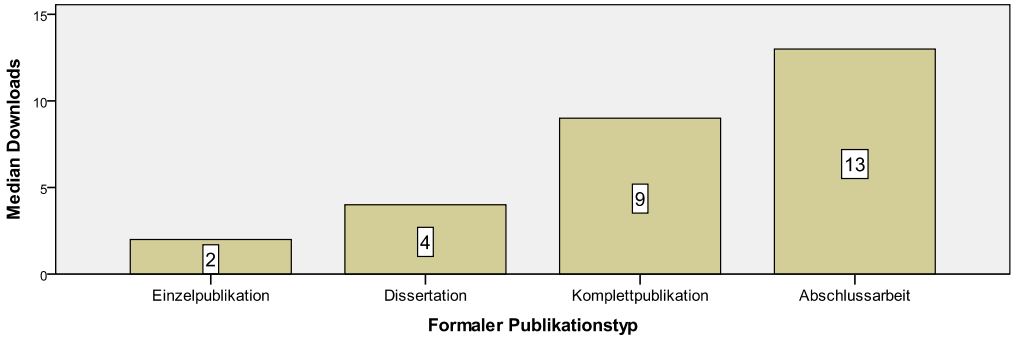


Abb. 31: Mediane der Downloads nach Publikationstyp, HeiDOK 7/2009

Da die Kategorien des Merkmals Publikationstyp beibehalten wurden, erübrigt sich eine Darstellung der Entwicklung des Dokumentenbestandes, da diese schon in Schritt 2 dargestellt wurde.

Fakultät von Dissertationen

Im Gegensatz zur Analyse des Publikationstyps mussten Kategorien bei der Auswertung der Fakultäten zusammengefasst werden, um zu signifikant unterschiedlichen Gruppen zu kommen. Für diese Gruppen, die durch die Variable F_Diss_Sign gebildet werden, sind die Mediane zu berechnen und grafisch darzustellen.

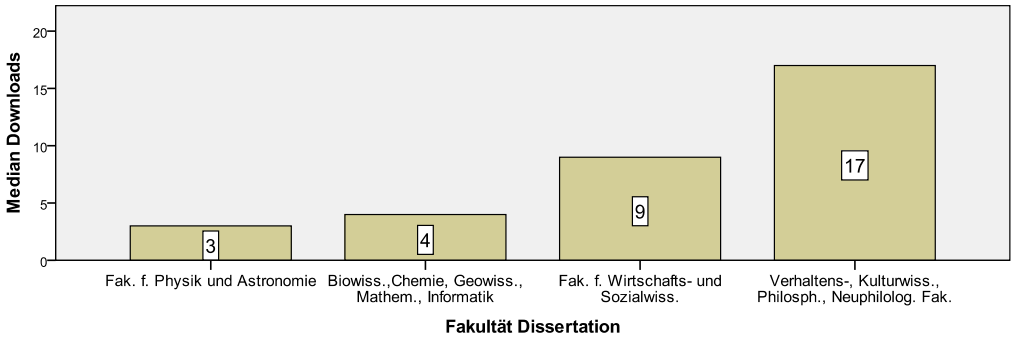


Abb. 32: Mediane der Downloads von Dissertationen nach Fakultät, HeiDOK 7/2009

Um die Entwicklung des Publikationsbestandes in diesen Gruppen darzustellen, wird vom aktuellen Metadatenfile ausgegangen, auf welches die gleiche Transformation der Variablen F_Diss nach F_Diss_Sign wie beim Downloadfile angewendet wird. Die folgende Abbildung zeigt die Entwicklung des Publikationsbestandes in den Gruppen der Variablen F_Diss_Sign, d.h. in den Gruppen, deren Mediane in der vorangegangenen Abbildung enthalten sind.

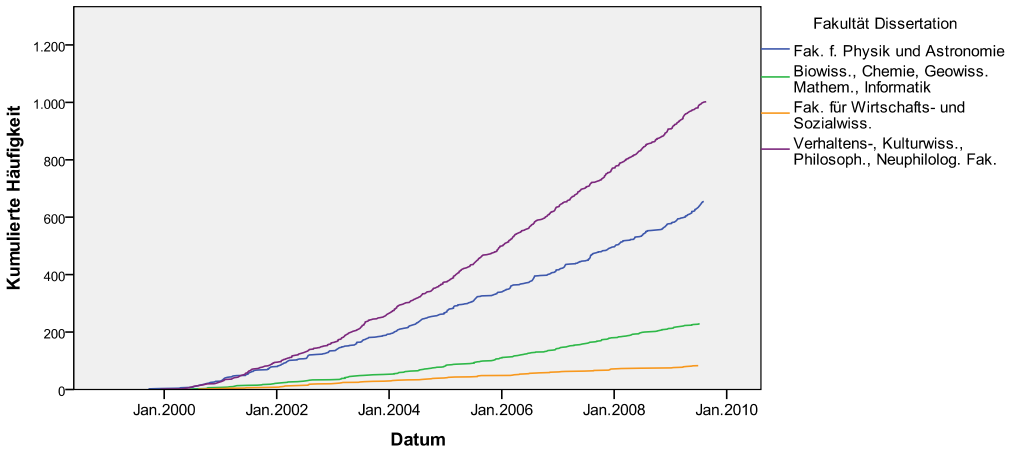


Abb. 33: Entwicklung des Publikationsbestandes an Dissertationen in den Ergebnisgruppen nach Fakultät, HeiDOK 7/2009

Mit Schritt 6 wurde die Analyse abgeschlossen. Die Ergebnisse können nun interpretiert werden.

5 Ergebnisse

Die Methode NoRA wurde mit einheitlichen Kategorien der Merkmale „Formaler Publikationstyp“ und „Inhaltsklasse“ und mit individuellen Kategorien des Merkmals „Fakultät“ zur Analyse aller vier universitärer Repositories angewendet, wobei an dieser Stelle die nachgelieferten Daten von 2010 für HeiDOK zugrunde liegen. Die folgenden Ergebnisse stellen nur eine Auswahl der möglichen einzelnen Analysen dar, die je nach den Bedürfnissen der Betreiber erweitert und verändert werden können. Hier sollen Ergebnisse aufgeführt werden, die auch Vergleiche der Repositories ermöglichen, wobei die absoluten Werte nur dann verglichen werden können, wenn es sich um die Analyse der Metadaten handelt. In den Grafiken zur Entwicklung des Bestandes an Publikationen sind nur die Gruppen enthalten, die mindestens 1 % des Gesamtbestandes bilden und mindestens 10 Publikationen enthalten. Die bei der Analyse der Nutzungsdaten ausgeschlossenen Kategorien sind rot markiert. Auf eine detaillierte Analyse der Komplettpublikationen wurde verzichtet, weil ihr Anteil bei allen vier Repositories unter 10 % liegt.

Ein Vergleich der Ergebnisse der Nutzungsdatenanalysen ist nur insoweit sinnvoll, als dass die Reihenfolgen und damit Tendenzen miteinander verglichen werden. Mit dem Test von Kruskal-Wallis wird die Hypothese $H_{0,}$ dass die Verteilung der Variablen Downloads in den Kategorien von Publikationstyp, Inhaltsklasse oder Fakultät gleich ist, geprüft. Bei Ablehnung der Hypothese wird überprüft, welche paarweisen Vergleiche signifikant sind. Sollte die Hypothese H_0 nicht abgelehnt werden, wurden keine signifikanten Unterschiede festgestellt und es können keine Vergleiche der Downloads erfolgen.

In den Balkendiagrammen sind ausschließlich die Gruppen von Publikationen mit Angabe der Mediane enthalten, für die sich die Verteilungen der Downloads paarweise signifikant unterscheiden. Unterscheiden sich Gruppen signifikant und die Mediane sind gleich, ist in den Ergebnissen zusätzlich die Ausgabe des Kruskal-Wallis-Tests für die paarweisen Vergleiche angegeben worden, da anhand der enthaltenen Mittelwerte der Ränge die Reihenfolge in den Balkendiagrammen bestimmt wurde. In diesen Fällen sind die Unterschiede der Verteilungen der Downloads zwar signifikant, aber so gering, dass sie vernachlässigt werden können. Nachdem die signifikant unterschiedlichen Gruppen von Publikationen, genannt Ergebnisgruppen, ermittelt wurden, wird die Entwicklung des Bestandes dieser Gruppen grafisch dargestellt.

Für jedes IR wird im Anschluss an die konkreten Ergebnisse von NoRA eine Zusammenfassung gegeben. Im Fazit wird auf ausgewählte spezifische Probleme eingegangen. Wenn möglich, werden Vorschläge gemacht, wie versucht werden kann, die Qualität des IR zu verbessern. Konkretere Maßnahmen können nur vom Betreiber selbst erwogen werden. Auf zwei Punkte soll bereits an dieser Stelle hingewiesen werden:

- Die Kategorien der untersuchten Merkmale mit den höchsten Downloads sind in der Regel zahlenmäßig gering vertreten. Deshalb kann durch verstärkte Akquisition von Publikationen auf diesen Gebieten die Sichtbarkeit der IR verbessert werden.

- Wie aus den folgenden Häufigkeitstabellen hervorgeht, gibt es bei allen IR Inhaltsklassen und Fakultäten, die wegen der geringen Fallzahlen oder weil sie gar nicht erst vorhanden sind, nicht in die Analyse der Downloads eingehen. Möglicherweise verbergen sich hier unbekannte Potentiale, die erschlossen werden sollten.

Diese Aussagen treffen auf alle vier IR zu und werden deshalb in den einzelnen Fazits nicht wiederholt.

5.1 Ergebnisse von edoc (Berlin)

5.1.1 Analyse der Metadaten

Analysiert werden 7049 Publikationen mit Volltexten im PDF-Format.

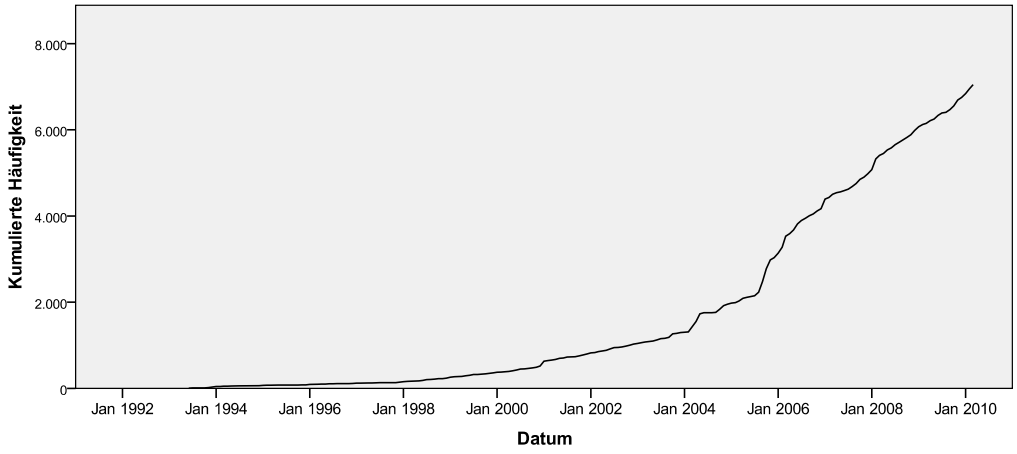


Abb. 34: Entwicklung des Bestandes an Publikationen, edoc 3/2010

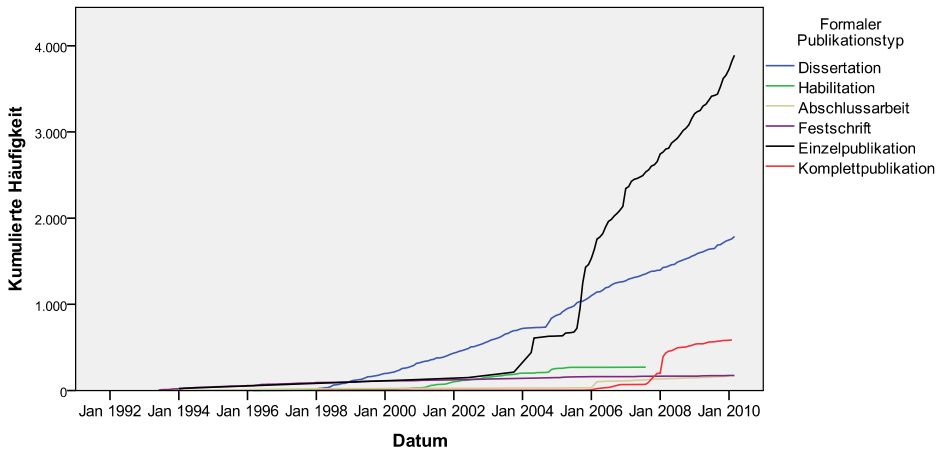


Abb. 35: Entwicklung des Bestandes an Publikationen nach Publikationstyp, edoc 3/2010

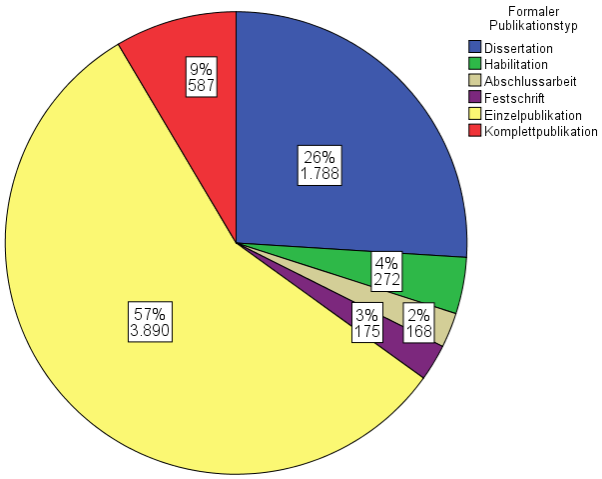


Abb. 36: Verteilung nach Publikationstyp, edoc 3/2010

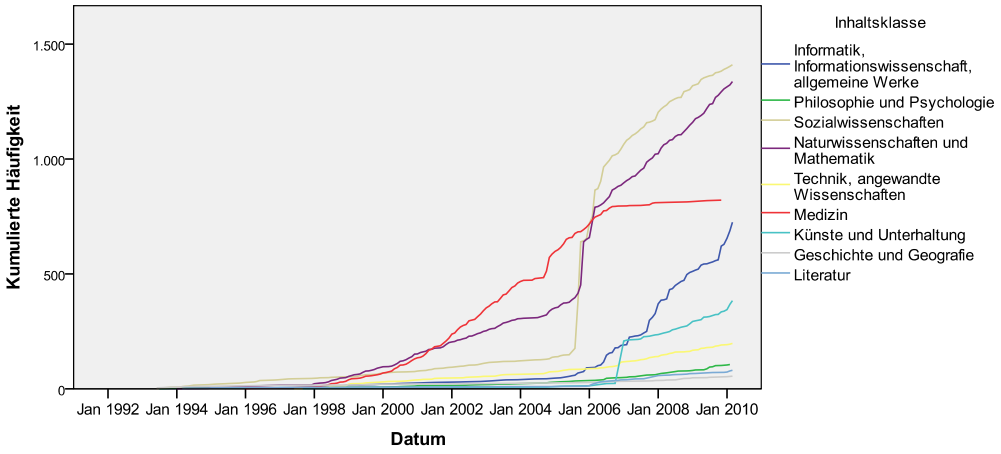


Abb. 37: Entwicklung des Bestandes an Publikationen nach Inhaltsklasse, edoc 3/2010

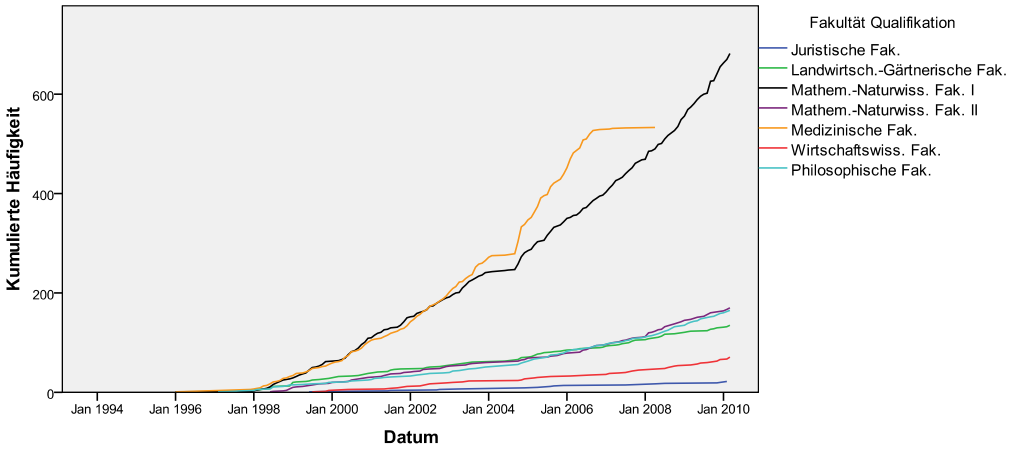


Abb. 38: Entwicklung des Bestandes an Dissertationen nach Fakultäten, edoc 3/2010

Wegen der Übersichtlichkeit wurden die Philosophischen Fakultäten zusammengefasst.

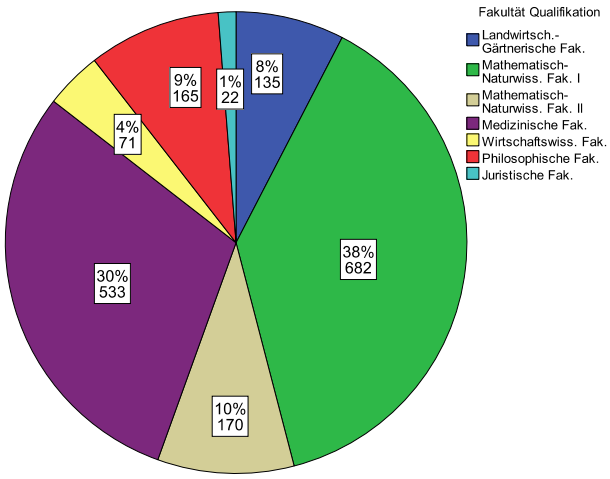


Abb. 39: Verteilung der Dissertationen nach Fakultät, edoc 3/2010

5.1.2 Analyse der Nutzungsdaten

Analysiert werden die Downloads der PDF-Files von 6150 Publikationen.

5.1.2.1 Formaler Publikationstyp

Tab. 28: Häufigkeiten Publikationstyp, edoc Downloadfile

P_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	9600	26,0	26,7	26,7
	B Habilitation	1632	4,4	4,5	31,3
	C Abschlussarbeit	900	2,4	2,5	33,8
	D Festschrift	1014	2,7	2,8	36,6
	E Einzelpublikation	19500	52,8	54,3	90,9
	F Komplettpublikation	3252	8,8	9,1	100,0
	Gesamt	35898	97,3	100,0	
Fehlend		1002	2,7		
Gesamt		36900	100,0		

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Kategorien sind signifikant. Fasst man A, B, und C zusammen, erhält man für alle paarweisen Vergleiche Signifikanz. Die Mediane für Komplett- und Einzelpublikationen sind gleich. Die Reihenfolge in der Grafik ergibt sich aus dem Vergleich der mittleren Ränge.

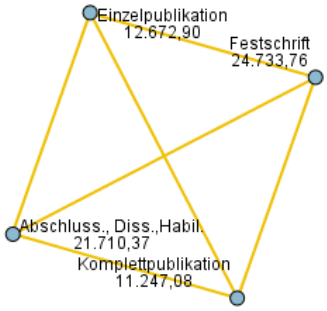


Abb. 40: Paarweise Vergleiche Publikationstyp, edoc 3/2010

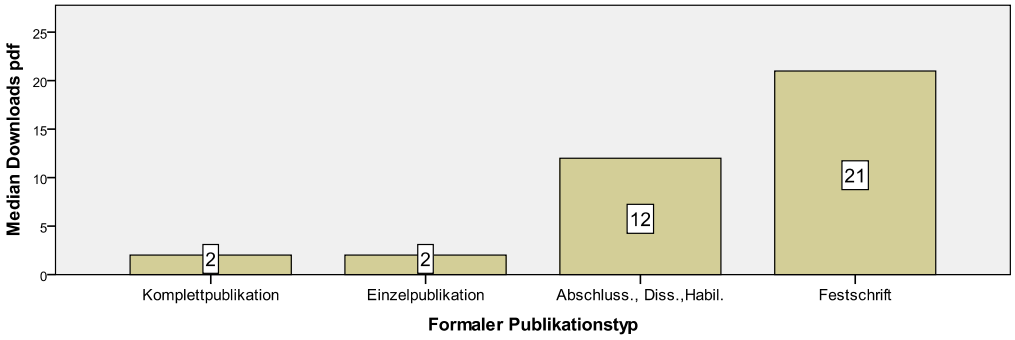


Abb. 41: Mediane der Downloads nach Publikationstyp, edoc 3/2010

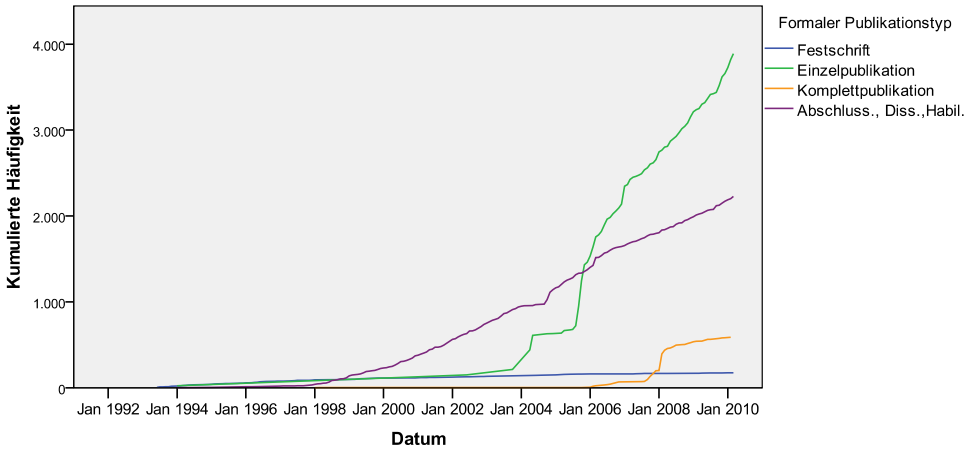


Abb. 42: Entwicklung des Bestandes in Ergebnisgruppen nach Publikationstyp, edoc 3/2010

5.1.2.2 Inhaltsklasse

Tab. 29: Häufigkeiten Inhaltsklasse, edoc Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	3126	8,5	11,4	11,4
	1 Philosophie und Psychologie	492	1,3	1,8	13,2

	2 Religion	114	,3	,4	13,6
	3 Sozialwissenschaften	7962	21,6	29,1	42,8
	4 Sprache	198	,5	,7	43,5
	5 Naturwissenschaften und Mathematik	7086	19,2	25,9	69,4
	6 Technik, angewandte Wissenschaften	1020	2,8	3,7	73,1
	61 Medizin	4884	13,2	17,9	91,0
	7 Künste und Unterhaltung	1776	4,8	6,5	97,5
	8 Literatur	402	1,1	1,5	98,9
	9 Geschichte und Geografie	288	,8	1,1	100,0
	Gesamt	27348	74,1	100,0	
Fehlend		9552	25,9		
Gesamt		36900	100,0		

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Kategorien sind signifikant. Fasst man 1, 6, und 8 zusammen, ergibt sich Signifikanz für alle paarweisen Vergleiche.

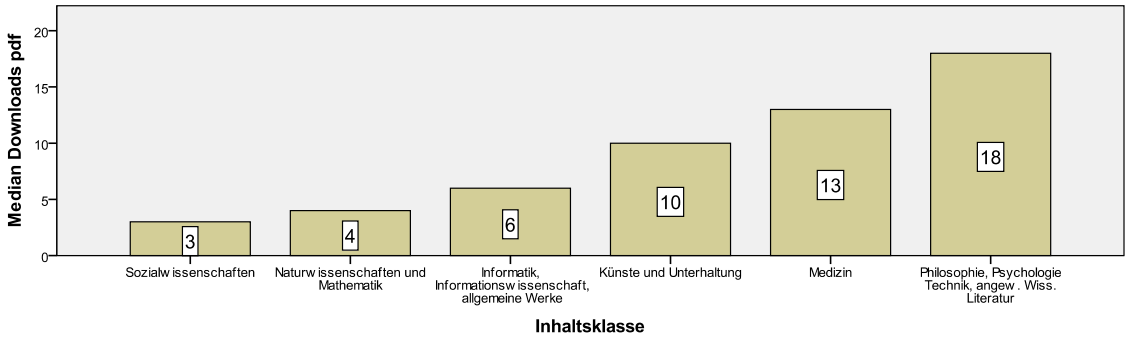


Abb. 43: Mediane der Downloads nach Inhaltsklasse, edoc 3/2010

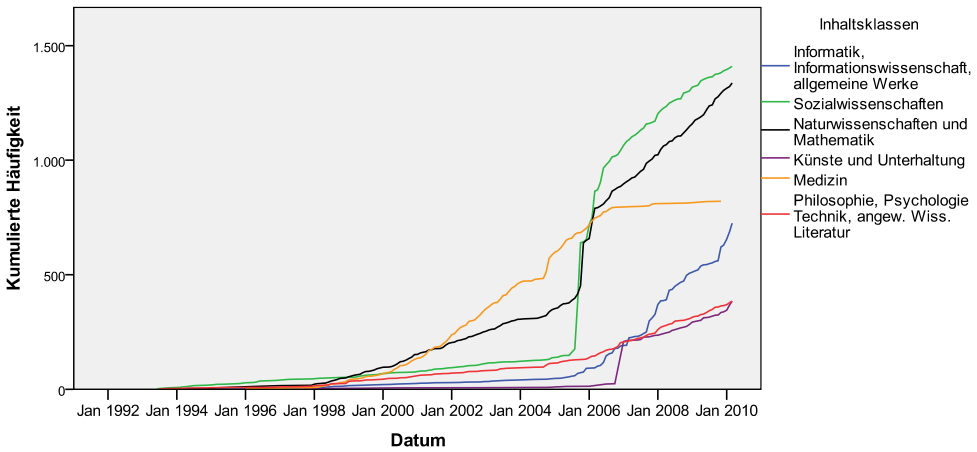


Abb. 44: Entwicklung des Bestandes in Ergebnisgruppen nach Inhaltsklasse, edoc 3/2010

5.1.2.3 Inhaltsklasse von Einzelpublikationen

Tab. 30: Häufigkeiten Inhaltsklasse von Einzelpublikationen, edoc Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	2214	11,4	16,7	16,7
	1 Philosophie und Psychologie	66	,3	,5	17,2
	2 Religion	42	,2	,3	17,5
	3 Sozialwissenschaften	5988	30,7	45,1	62,6
	4 Sprache	36	,2	,3	62,8
	5 Naturwissenschaften und Mathematik	2772	14,2	20,9	83,7
	6 Technik, angewandte Wissenschaften	216	1,1	1,6	85,3
	7 Künste und Unterhaltung	1632	8,4	12,3	97,6
	8 Literatur	288	1,5	2,2	99,8
	9 Geschichte und Geografie	30	,2	,2	100,0
	Gesamt	13284	68,1	100,0	
Fehlend		6216	31,9		
Gesamt		19500	100,0		

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Kategorien sind signifikant. Fasst man 6 und 8 zusammen, erhält man Signifikanz bei allen paarweisen Vergleichen.

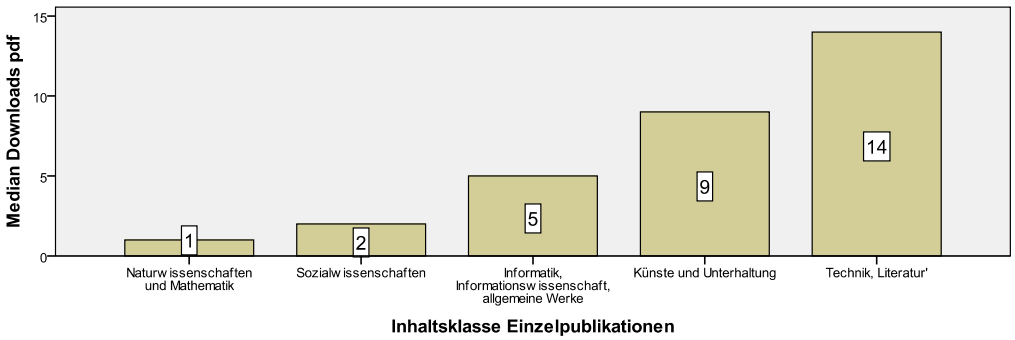


Abb. 45: Mediane der Downloads von Einzelpublikationen nach Inhaltsklasse, edoc 3/2010

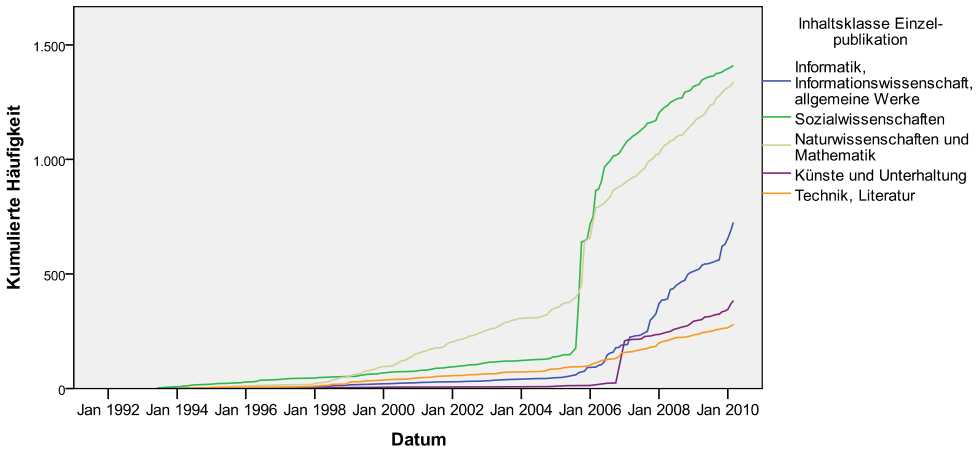


Abb. 46: Entwicklung des Bestandes an Einzelpublikationen in Ergebnisgruppen nach Inhaltsklasse, edoc 3/2010

5.1.2.4 Fakultät

Die Auswertung nach Fakultäten wird nur für die Gruppe „Abschlussarbeit, Dissertation, Habilitation“ (Qualifikation) durchgeführt, da für andere Publikationen Fakultäten selten angegeben sind. Beim Typ „Festschrift“ handelt es sich nur um Publikationen der Philosophischen Fakultäten. Im Unterschied zu den anderen IR werden nicht die Dissertationen allein, sondern mit Habilitationen und Abschlussarbeiten zusammen analysiert, da diese Publikationstypen sich hinsichtlich der Verteilung der Downloads nicht unterscheiden.

Tab. 31: Häufigkeiten Fakultät Qualifikation, edoc Downloadfile

D_Fak Fakultät Qualifikation					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 Juristische Fak.	108	,9	,9	,9
	2 Landwirtsch. -Gärtnerische Fak.	750	6,2	6,2	7,1
	3 Mathematisch-Naturwiss. Fak. I	3600	29,7	29,7	36,7
	4 Mathematisch-Naturwiss. Fak. II	948	7,8	7,8	44,6
	5 Medizinische Fak.	4770	39,3	39,3	83,9
	6 Philosophische Fak. I	264	2,2	2,2	86,1
	7 Philosophische Fak. II	126	1,0	1,0	87,1
	8 Philosophische Fak. III	390	3,2	3,2	90,3
	9 Philosophische Fak. IV	168	1,4	1,4	91,7
	10 Theologische Fak.	18	,1	,1	91,8
	11 Wirtschaftswiss. Fak.	954	7,9	7,9	99,7
	12 Missing	36	,3	,3	100,0
	Gesamt	12132	100,0	100,0	

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Kategorien sind signifikant. Um bei allen paarweisen Vergleichen Signifikanz zu erzeugen, wurden alle Philosophischen Fakultäten und 5 und 11 zusammengefasst.

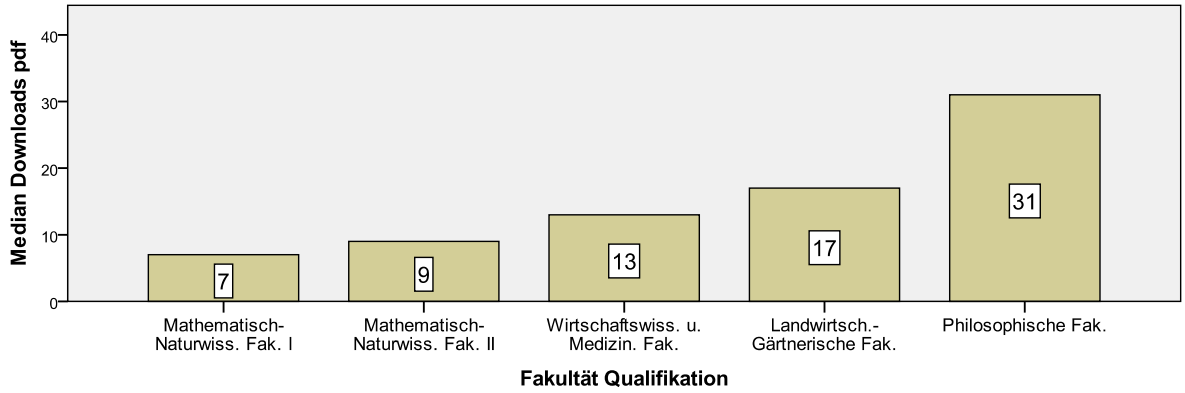


Abb. 47: Mediane der Downloads von Qualifikationsarbeiten nach Fakultät, edoc 3/2010

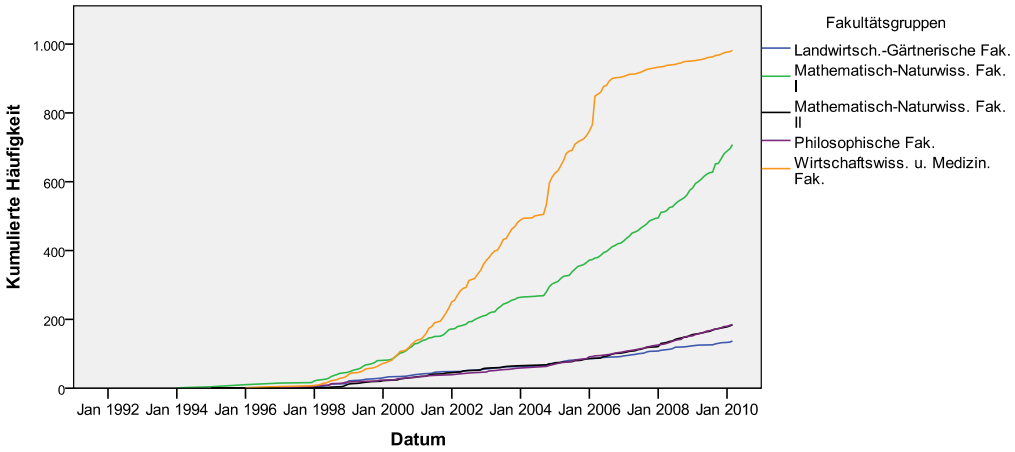


Abb. 48: Entwicklung des Bestandes an Qualifikationsarbeiten in Ergebnisgruppen nach Fakultät, edoc 3/2010

5.1.3 Zusammenfassung

Formaler Publikationstyp

Die meisten Zugriffe gibt es auf Publikationen vom Typ „Festschrift“, gefolgt von den Qualifikationsarbeiten. Die Downloads von Komplett- und Einzelpublikationen befinden sich im untersten Bereich. Den stärksten Anstieg der Anzahl von Publikationen gibt es bei Einzelpublikationen, also bei den Publikationen mit den wenigsten Downloads. Die Anzahl der Qualifikationsarbeiten, an denen Dissertationen den überwiegenden Anteil bilden, vergrößert sich gleichmäßig.

Inhaltsklasse

Die meisten Downloads (Median 18) gibt es bei Publikationen aus den Klassen „Religion“, „Technik und angewandte Wissenschaften“ und „Literatur“. Die Anzahl an Publikationen steigt gleichmäßig, wenn auch auf niedrigem Niveau. Im mittleren Bereich (Mediane 6-13) stagniert die Klasse „Medizin“, die Klasse „Künste und Unterhaltung“ nimmt in der Anzahl gleichmäßig auf niedrigem Niveau zu, „Informatik, Informationswissenschaften und allgemeine Werke“ verzeichnen den steilsten Anstieg. Die Klassen mit den wenigsten Downloads (Mediane 1 und 4) sind „Sozialwissenschaften“ und „Naturwissenschaften und Mathematik“, die den größten Anteil an Publikationen darstellen und deren Anzahl gleichmäßig wächst. Betrachtet man die Einzelpublikationen extra, befinden sich die gleichen Klassen im untersten Bereich der Downloads.

Fakultät von Qualifikationsarbeiten

Die mit Abstand meisten Downloads (Median 31) gibt es bei den Qualifikationsarbeiten der Philosophischen Fakultäten. Deren Anteile steigen auf niedrigem Niveau. Im mittleren Bereich befinden sich die Wirtschaftswissenschaftliche und die Medizinische Fakultät, deren Anteile an den gesamten Qualifikationsarbeiten am größten sind. Grund dafür ist der hohe Anteil an medizinischen Dissertationen, deren Anteil nicht mehr wächst, da keine neuen medizinischen Dissertationen veröffentlicht werden. Die Anzahl der Dissertationen aus der Landwirtschaftlich-Gärtnerischen Fakultät wächst gleichmäßig auf niedrigem Niveau. Die Qualifikationsarbeiten der Mathematisch-Naturwissenschaftlichen Fakultäten befinden sich im Bereich der niedrigsten Downloads (Median 7 und 9), wobei die Anzahl der Qualifikationsarbeiten aus dem niedrigsten Bereich am stärksten überhaupt wächst.

Fazit

Die Anzahl der Einzelpublikationen wächst seit Jahren kontinuierlich am stärksten und bildet den größten Anteil an den Publikationen. Damit wächst gerade der Publikationstyp am schnellsten, dessen Downloads sich im niedrigsten Bereich befinden. Es müssen Maßnahmen getroffen werden, um die Sichtbarkeit der Einzelpublikationen zu erhöhen. Naheliegend ist, die Homepage von edoc so umzugestalten, dass die Anzahl der Ebenen, über die der Nutzer zur Einzelpublikation gelangt, verringert wird.

Ein ähnliches Bild ergibt sich, wenn man die Downloads und die Entwicklung nach Inhaltsklassen betrachtet. Die Klassen mit dem größten Anteil und stärkstem Zuwachs, „Sozialwissenschaften“ und „Naturwissenschaften und Mathematik“, haben die niedrigsten Downloads. Eine Ausnahme bilden die Qualifikationsarbeiten, bei denen ein solcher Trend nicht generell beobachtet wird. Allerdings haben auch hier die Publikationen mit geringem Anteil an der Gesamtanzahl der Gruppe die höchsten Downloads. Eine Maßnahme, dem Trend entgegenzuwirken, könnte die Auswahlmöglichkeit von Inhaltsklassen oder Fakultäten auf einer oberen Ebene sein, um den Nutzer geradlinig zum Fachgebiet zu führen.

5.2 Ergebnisse von ehsStu (Stuttgart)

Bei der Analyse von ehsStu wird vom Prinzip der für alle IR gemeinsam verwendeten Publikationstypen abgewichen und die Kategorie „Studienarbeit“, die nur hier zu finden ist, berücksichtigt. Ebenso werden, da keine monatlichen Downloads vorhanden sind, andere Erhebungszeiträume verwendet.

5.2.1 Analyse der Metadaten

In die Analyse gehen die Metadaten von 4215 Publikationen ein.

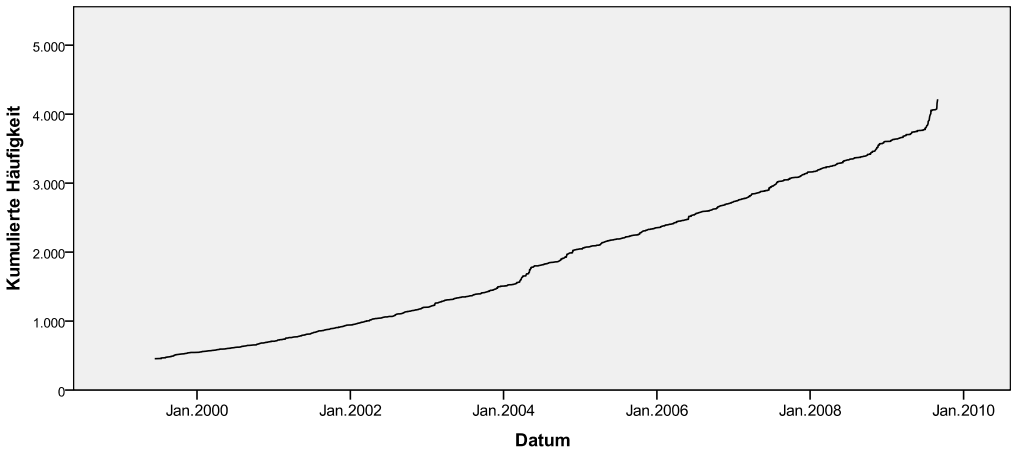


Abb. 49: Entwicklung des Bestandes an Publikationen, ehsStu 8/2009

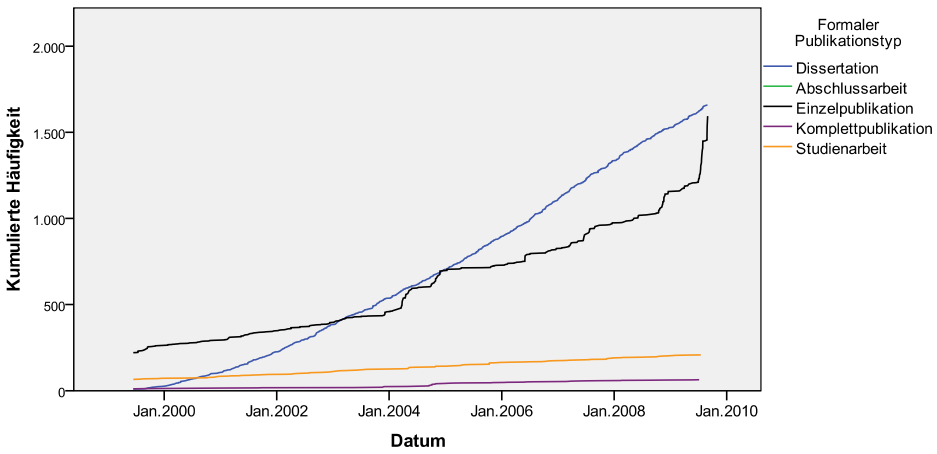


Abb. 50: Entwicklung des Bestandes an Publikationen nach Publikationstyp, ehsStu 8/2009

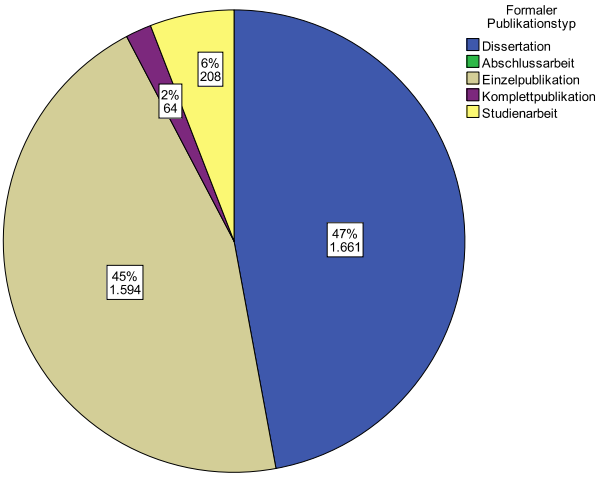


Abb. 51: Verteilung des Publikationstyps, ehsStu 8/2009

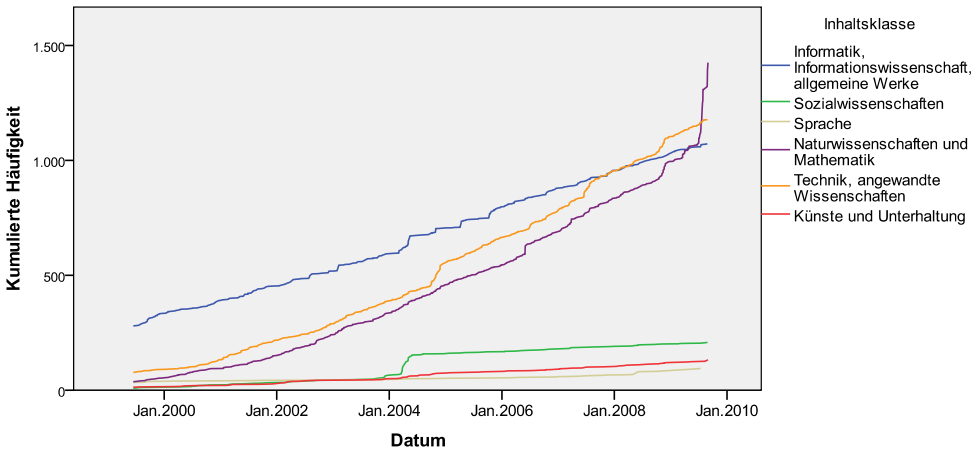


Abb. 52: Entwicklung des Bestandes an Publikationen nach Inhaltsklasse, ehsStu 8/2009

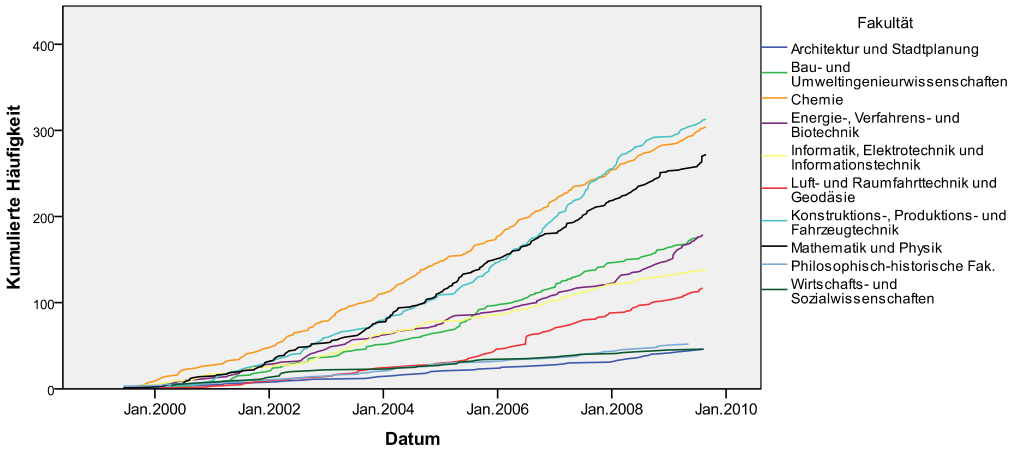


Abb. 53: Entwicklung des Bestandes an Dissertationen nach Fakultäten, ehsStu 8/2009

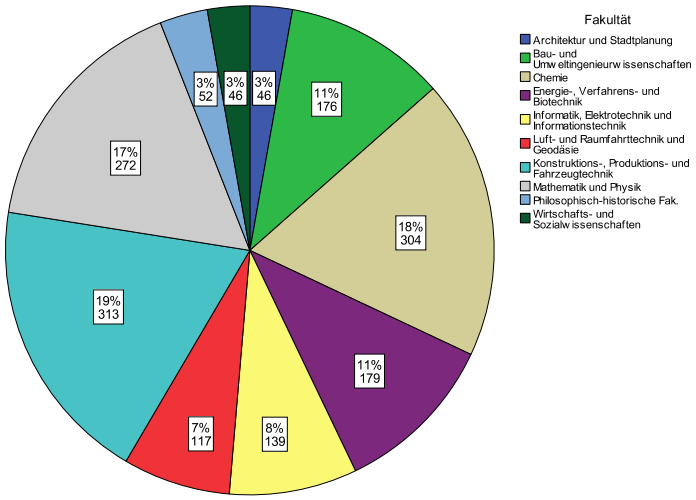


Abb. 54: Verteilung von Dissertationen nach Fakultät, ehsStu 8/2009

5.2.2 Analyse der Nutzungsdaten

In die Analyse gehen die Downloads von PDF-Files von 2545 Publikationen ein.

Im Unterschied zu den anderen drei IR liegen für ehsStu nur Nutzungsdaten für drei Erhebungszeiträume und damit für jede Publikation nur halb so viele Downloads vor. Die Erhebungszeiträume sind 01-06/2007, 07-12/2007 und 11-12/2008 und damit nicht zusammenhängend. Die Downloads des letzten Erhebungszeitraumes wurden mit 3 multipliziert, um sie auswerten zu können. Damit unterscheiden sich die Downloads von ehsStu von der Erhebung her deutlich von denen der anderen IR. Trotzdem konnte die Methode mit signifikanten Ergebnissen auf diese Daten angewendet werden.

5.2.2.1 Formaler Publikationstyp

Tab. 32: Häufigkeiten Publikationstyp, ehsStu Downloadfile

P_S_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	2970	38,9	49,0	49,0
	B Habilitation	57	,7	,9	49,9
	D Festschrift	9	,1	,1	50,0
	E Einzelpublikation	2373	31,1	39,1	89,2
	F Komplettpublikation	153	2,0	2,5	91,7
	S Studienarbeit	504	6,6	8,3	100,0
	Gesamt	6066	79,4	100,0	
Fehlend		1569	20,6		
Gesamt		7635	100,0		

Die Hypothese H_0 wird abgelehnt. Fasst man Einzelpublikationen und Studienarbeiten zusammen, sind alle paarweisen Vergleiche signifikant.

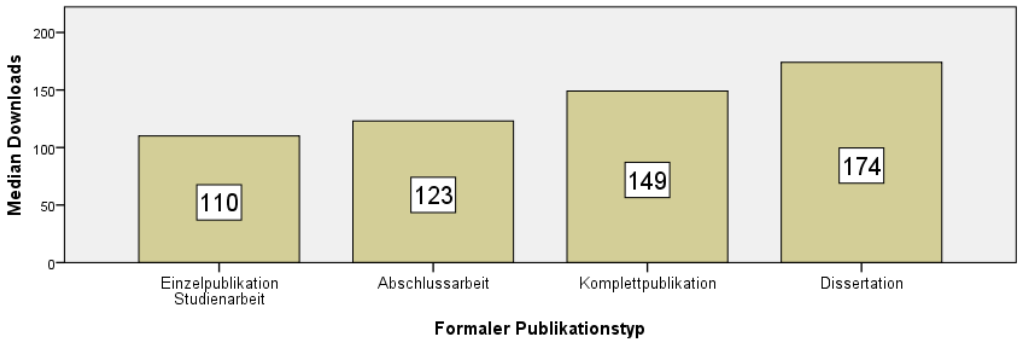


Abb. 55: Mediane der Downloads nach Publikationstyp, ehsStu 12/2008

5.2.2.2 Inhaltsklasse

Tab. 33: Häufigkeiten Inhaltsklasse, ehsStu Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	2514	32,9	32,9	32,9
	1 Philosophie und Psychologie	69	,9	,9	33,8
	3 Sozialwissenschaften	519	6,8	6,8	40,6
	4 Sprache	168	2,2	2,2	42,8
	5 Naturwissenschaften und Mathematik	1908	25,0	25,0	67,8
	6 Technik, angewandte Wissenschaften	2130	27,9	27,9	95,7
	61 Medizin	24	,3	,3	96,0
	7 Künste und Unterhaltung	255	3,3	3,3	99,4
	8 Literatur	15	,2	,2	99,6
	9 Geschichte und Geografie	33	,4	,4	100,0
	Gesamt	7635	100,0	100,0	

Die Hypothese H_0 wird abgelehnt. Werden 0 und 4 zusammengefasst, sind alle paarweisen Vergleiche signifikant.

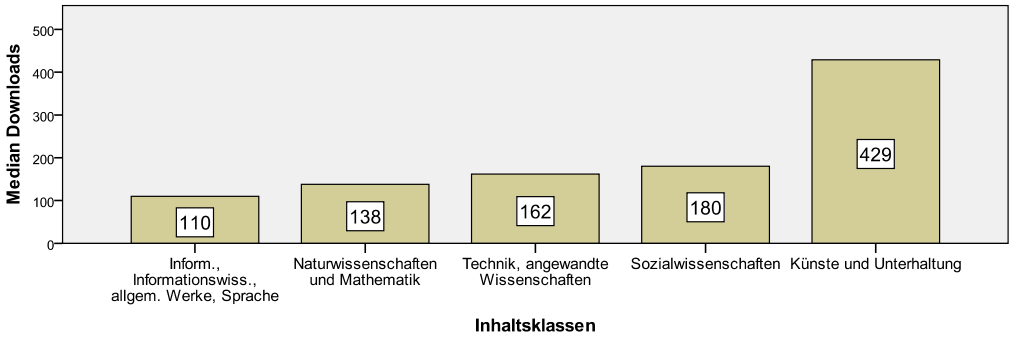


Abb. 56: Mediane der Downloads nach Inhaltsklasse, ehsStu 12/2008

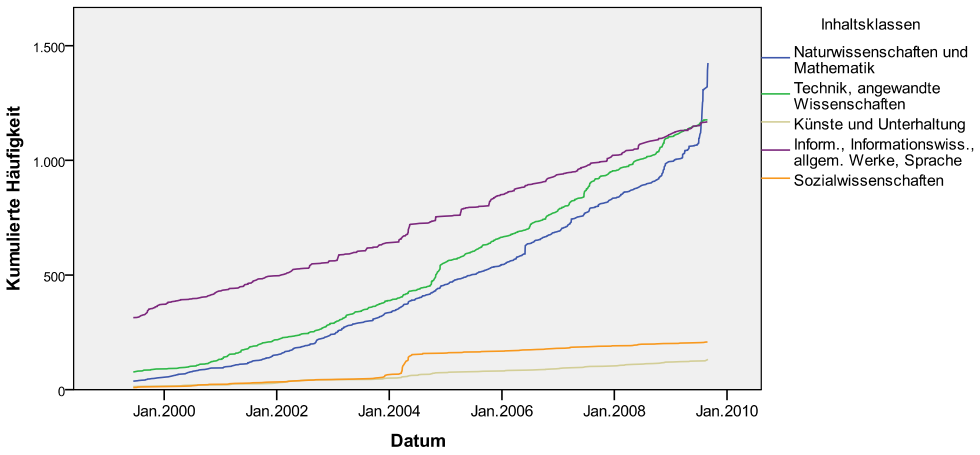


Abb. 57: Entwicklung des Bestandes in Ergebnisgruppen nach Inhaltsklasse, ehsStu 8/2009

5.2.2.3 Inhaltsklasse von Einzelpublikationen

Tab. 34: Häufigkeiten Inhaltsklasse von Einzelpublikationen, ehsStu Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	708	29,8	29,8	29,8
	1 Philosophie und Psychologie	27	1,1	1,1	31,0
	3 Sozialwissenschaften	318	13,4	13,4	44,4
	4 Sprache	90	3,8	3,8	48,2
	5 Naturwissenschaften und Mathematik	318	13,4	13,4	61,6
	6 Technik, angewandte Wissenschaften	777	32,7	32,7	94,3
	61 Medizin	21	,9	,9	95,2
	7 Künste und Unterhaltung	96	4,0	4,0	99,2
	8 Literatur	6	,3	,3	99,5

	9 Geschichte und Geografie	12	,5	,5	100,0
	Gesamt	2373	100,0	100,0	

Die Hypothese H_0 wird abgelehnt. Signifikante Ergebnisse bei allen paarweisen Vergleichen erhält man erst, wenn man 0, 4, 5 und 6 zu einer Gruppe und 3 und 7 zu einer weiteren Gruppe zusammenfasst.

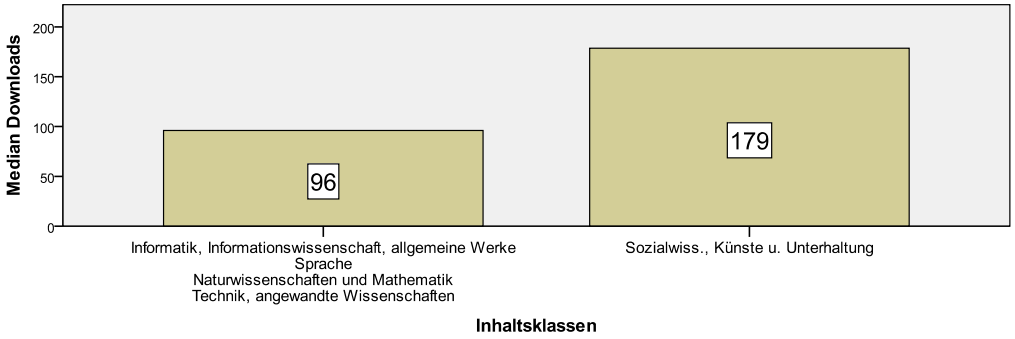


Abb. 58: Mediane der Downloads von Einzelpublikationen nach Inhaltsklasse, ehsStu, 12/2008

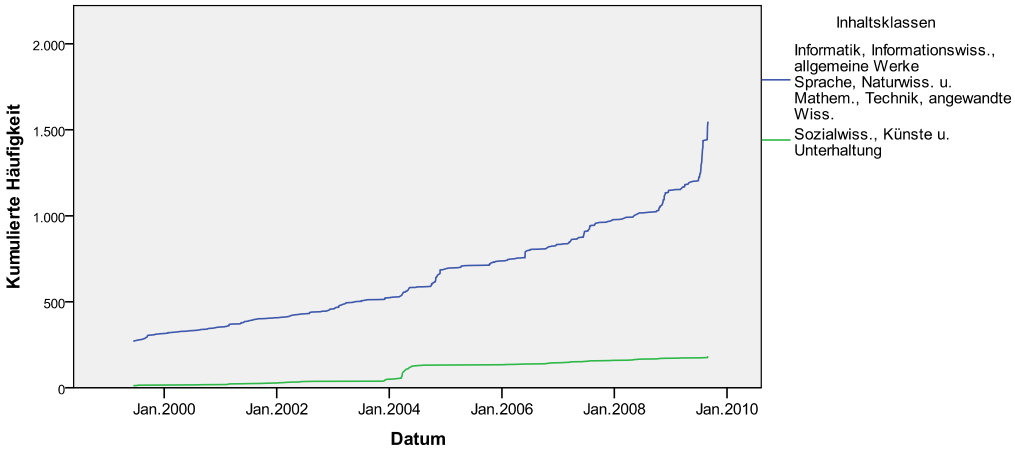


Abb. 59: Entwicklung des Bestandes an Einzelpublikationen in Ergebnisgruppen nach Inhaltsklasse, ehsStu 8/2009

5.2.2.4 Fakultät von Dissertationen

Tab. 35: Häufigkeiten Fakultät von Dissertationen, ehsStu Downloadfile

Fakultaet Fakultät					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	01 Architektur und Stadtplanung	78	2,6	2,7	2,7
	02 Bau- und Umweltingenieurwissenschaften	315	10,6	10,7	13,4
	03 Chemie	594	20,0	20,2	33,6
	04 Energie-, Verfahrens- und Biotechnik	291	9,8	9,9	43,5
	05 Informatik, Elektrotechnik und Informationstechnik	276	9,3	9,4	52,9
	06 Luft- und Raumfahrttechnik und Geodäsie	171	5,8	5,8	58,7
	07 Konstruktions-, Produktions- und Fahrzeugtechnik	501	16,9	17,1	75,8
	08 Mathematik und Physik	495	16,7	16,9	92,6
	09 Philosophisch-historische Fak.	102	3,4	3,5	96,1
	10 Wirtschafts- und Sozialwissenschaften	102	3,4	3,5	99,6
	17 Fakultätsübergreifend / Sonst. Einrichtung	9	,3	,3	99,9
	18 Deutsches Zentrum für Luft- und Raumfahrt (DLR)	3	,1	,1	100,0
	Gesamt	2937	98,9	100,0	
Fehlend		33	1,1		
Gesamt		2970	100,0		

Die Hypothese H_0 wird abgelehnt. Von den 45 paarweisen Vergleichen sind 11 nicht signifikant. Fasst man dagegen die Fakultäten unter sachlogischen Gesichtspunkten zusammen, erhält man Signifikanz für alle Gruppen von Fakultäten.

Tab. 36: Häufigkeiten Zusammenfassung von Fakultäten, ehsStu Downloadfile

Fak_Sign Fakultäten					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	G1 Architektur, Ingenieurwiss., Technik	1632	55,2	55,2	55,2
	G2 Chemie, Mathematik, Physik	1089	36,8	36,8	92,0
	G3 Philos.-histor. Fak., Wirtschafts- u. Sozialwiss.	237	8,0	8,0	100,0
	Gesamt	2958	100,0	100,0	

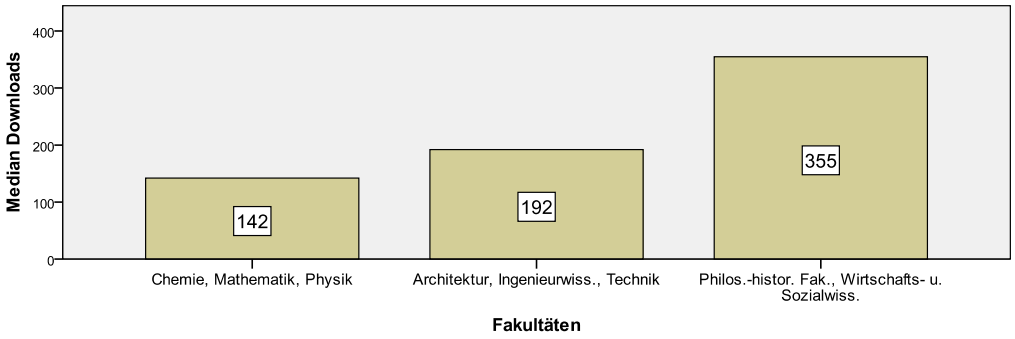


Abb. 60: Mediane der Downloads von Dissertationen nach Fakultät, ehsStu 12/2008

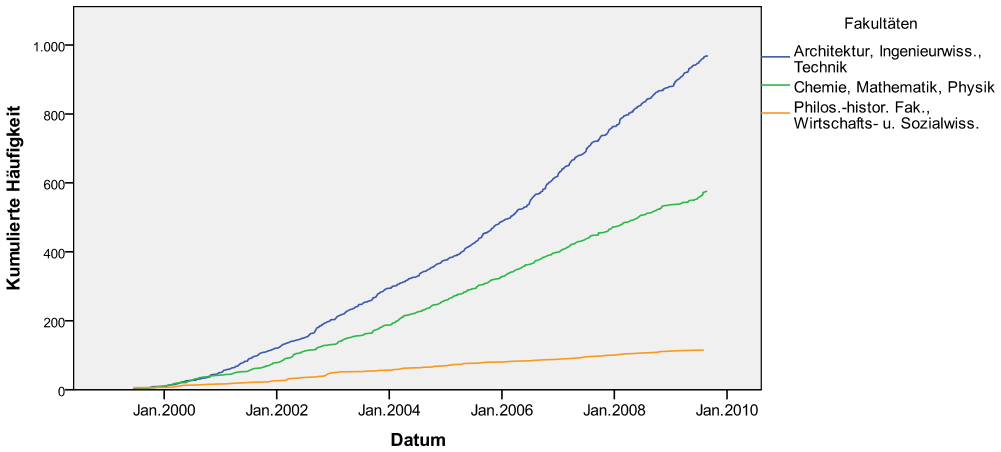


Abb. 61: Entwicklung des Bestandes an Dissertationen in Ergebnisgruppen nach Fakultätsgruppen, ehsStu 12/2008

5.2.3 Zusammenfassung

Formaler Publikationstyp

Dissertationen und Einzelpublikationen bilden mit 47 % bzw. 45 % der ausgewerteten Publikationen den größten Anteil an allen Publikationen. Dabei wächst die Anzahl von Einzelpublikationen schneller als die der anderen Kategorien. Dissertationen haben die meisten Downloads, Einzelpublikationen die wenigsten. Der Anteil der Abschlussarbeiten ist, verglichen mit den anderen IR, relativ hoch. Ihre Downloads sind höher als die von Einzelpublikationen, aber wesentlich geringer als die von Dissertationen.

Inhaltsklasse

Die Kategorie „Künste und Unterhaltung“ hat weitaus die höchsten Downloads. Wegen des erheblichen Unterschiedes zu den anderen Inhaltsklassen wurde ermittelt, welche Publikationstypen in der Klasse enthalten sind. Den größten Anteil bilden Dissertationen und Komplettpublikationen. Die hohen Downloads können deshalb durch Zugehörigkeit zu den beiden Publikationstypen und der Kategorie „Künste und Unterhaltung“ erklärt werden. Alle anderen analysierbaren Inhaltsklassen gehören in den unteren Bereich der Downloads. So auch die Kategorie „Naturwissenschaften und Mathematik“, die sich zahlenmäßig am stärksten entwickelt

und inzwischen den größten Anteil stellt. Bei den Einzelpublikationen konnten nur zwei Gruppen der Kategorien von „Inhaltsklasse“, die sich bezüglich der Downloads unterscheiden, identifiziert weden. Zur oberen Gruppe gehören „Sozialwissenschaften“ und „Künste und Unterhaltung“.

Fakultät

Ausgehend von den existierenden Fakultäten konnten keine Gruppen von Fakultäten ermittelt werden, die sich in ihren Downloads signifikant unterscheiden. Die Zusammenfassung unter sachlogischem Aspekt erwies sich als erfolgreich. Die Dissertationen der Philosophisch-historischen Fakultät und der Wirtschafts- und Sozialwissenschaften haben wesentlich mehr Downloads als die der anderen Fakultäten, ihr Anteil an der Gesamtzahl ist gering. Den größten Anteil haben die Dissertationen aus den Fakultäten der Bereiche Architektur, Ingenieurwissenschaften und Technik mit mittleren Downloads. Der Anteil von Disserationen aus Chemie, Mathematik und Physik liegt zwischen denen der beiden bereits genannten Gruppen. Ihre Downloads sind am niedrigsten.

Fazit

Es gibt bei einigen Kategorien starke Anstiege bei der Anzahl von Publikationen, so bei Inhaltsklasse „Naturwissenschaften und Mathematik“ und Einzelpublikationen. Bei Einzelpublikationen betrifft das die Zusammenfassung mehrerer Kategorien von „Inhaltsklasse“. Verglichen mit den Downloads anderer Publikationstypen sind die der Einzelpublikationen am niedrigsten. Ebenso liegen die Downloads der Inhaltsklasse „Naturwissenschaften und Mathematik“, verglichen mit anderen Kategorien, im unteren Bereich.

5.3 Ergebnisse von HeiDOK 2010 (Heidelberg)

5.3.1 Analyse der Metadaten

Analysiert werden die Metadaten von 3289 Publikationen.

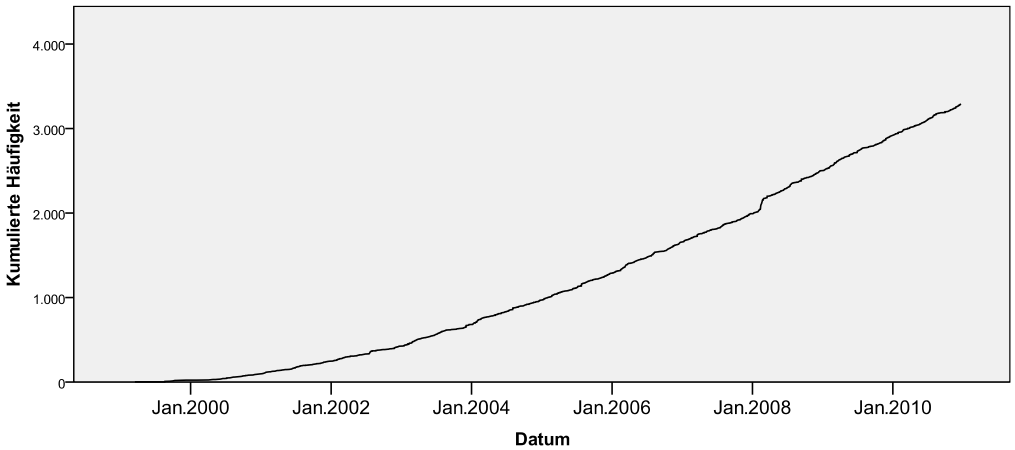


Abb. 62: Entwicklung des Bestandes an Publikationen, HeiDOK 12/2010

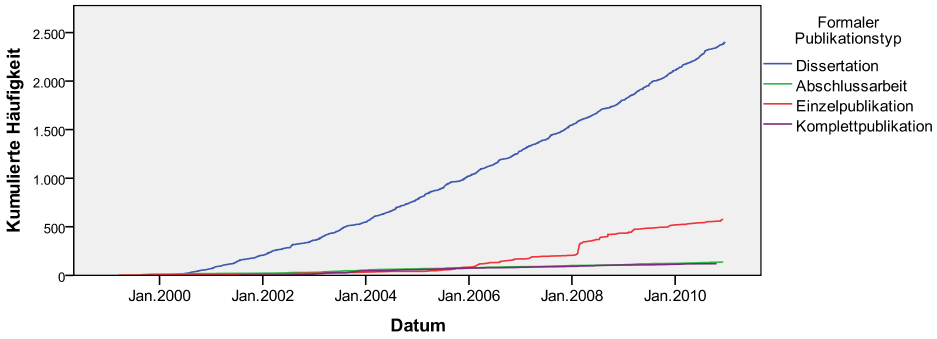


Abb. 63: Entwicklung des Bestandes an Publikationen nach Publikationstyp, HeiDOK 12/2010

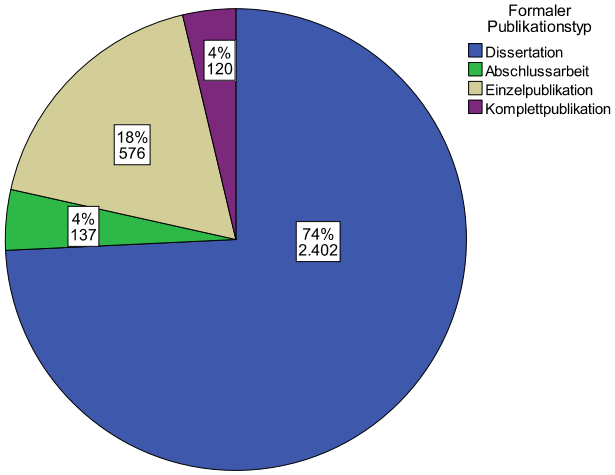


Abb. 64: Verteilung der Publikationen nach Publikationstyp, HeiDOK 12/2010

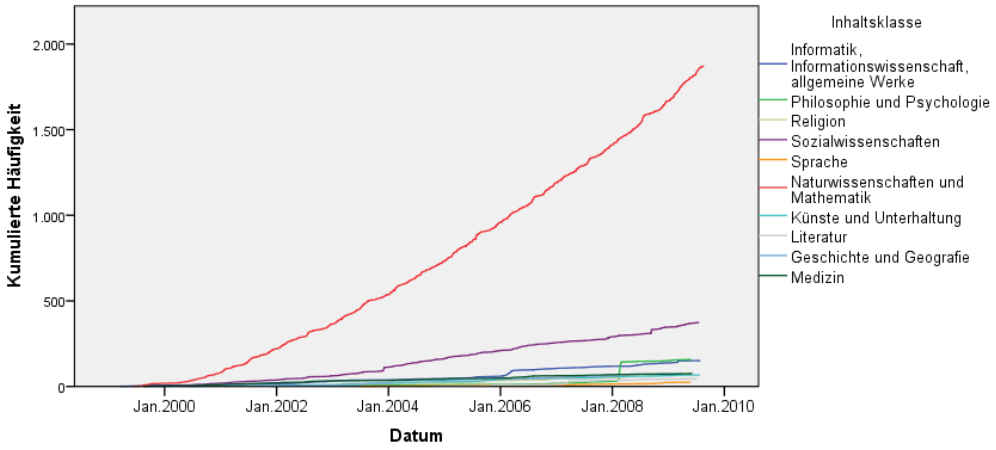


Abb. 65: Entwicklung des Bestandes an Publikationen nach Inhaltsklasse, HeiDOK 12/2010

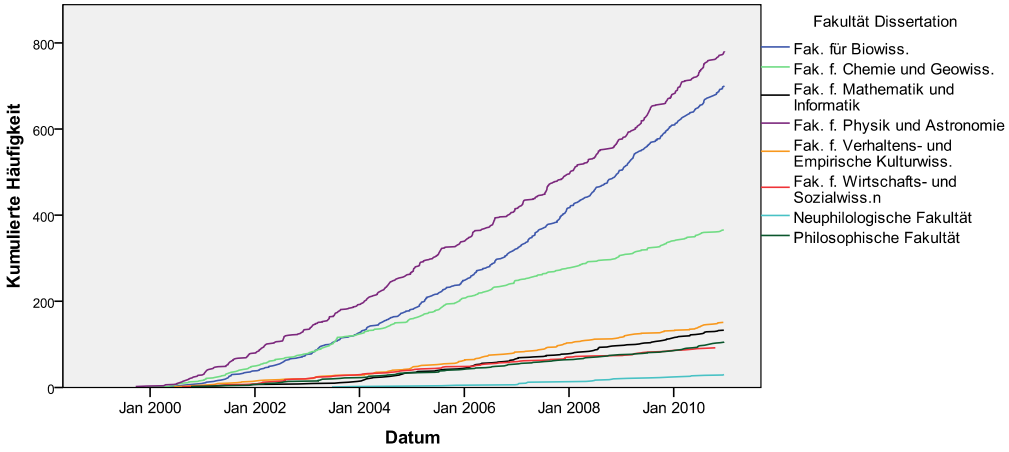


Abb. 66: Entwicklung des Bestandes an Dissertationen nach Fakultät, HeiDOK 12/2010

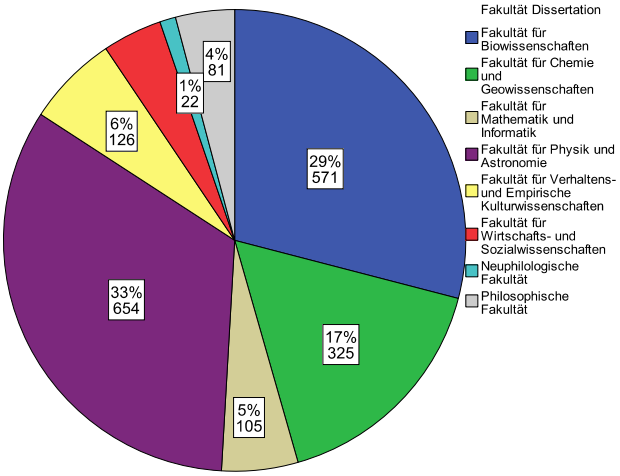


Abb. 67: Verteilung der Dissertationen nach Fakultät, HeiDOK 12/2010

5.3.2 Analyse der Nutzungsdaten

Analysiert werden die Daten von 2865 Publikationen.

5.3.2.1 Formaler Publikationstyp

Tab. 37: Häufigkeiten Publikationstyp, Downloadfile HeiDok 2010

P_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	12630	73,5	73,5	73,5
	B Habilitation	30	,2	,2	73,6
	C Abschlussarbeit	732	4,3	4,3	77,9
	E Einzelpublikation	3114	18,1	18,1	96,0
	F Komplettpublikation	684	4,0	4,0	100,0
	Gesamt	17190	100,0	100,0	

Die Hypothese H_0 wird abgelehnt und alle paarweisen Vergleiche der Publikationstypen sind signifikant.

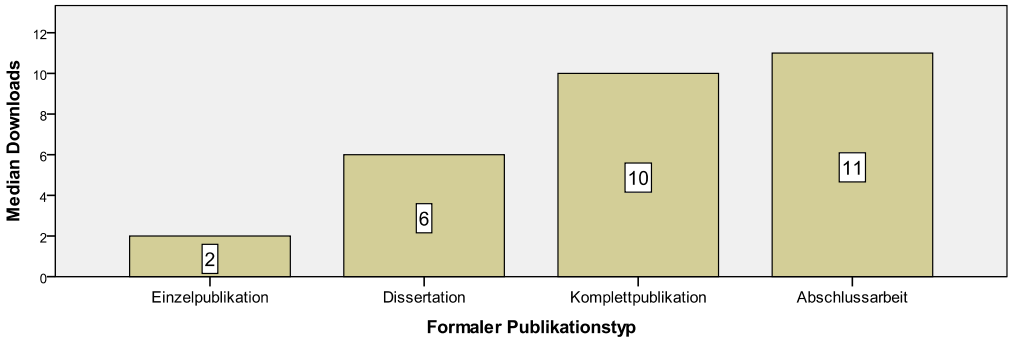


Abb. 68: Mediane der Downloads nach Publikationstyp, HeiDOK 12/2010

5.3.2.2 Inhaltsklasse

Tab. 38: Häufigkeiten Inhaltsklasse, HeiDOK 2010 Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	984	5,7	5,7	5,7
	1 Philosophie und Psychologie	972	5,7	5,7	11,4
	2 Religion	468	2,7	2,7	14,1
	3 Sozialwissenschaften	2124	12,4	12,4	26,5
	4 Sprache	114	,7	,7	27,1
	5 Naturwissenschaften und Mathematik	11052	64,3	64,3	91,4
	6 Technik, angewandte Wissenschaften	30	,2	,2	91,6
	61 Medizin	432	2,5	2,5	94,1
	7 Künste und Unterhaltung	390	2,3	2,3	96,4
	8 Literatur	282	1,6	1,6	98,0
	9 Geschichte und Geografie	342	2,0	2,0	100,0
	Gesamt	17190	100,0	100,0	

Die Hypothese H_0 wird abgelehnt. Fasst man die Kategorien 0 und 1 zu einer Gruppe und 2, 3, 61 und 9 zu einer weiteren Gruppe zusammen, enthält man signifikante paarweise Vergleiche.

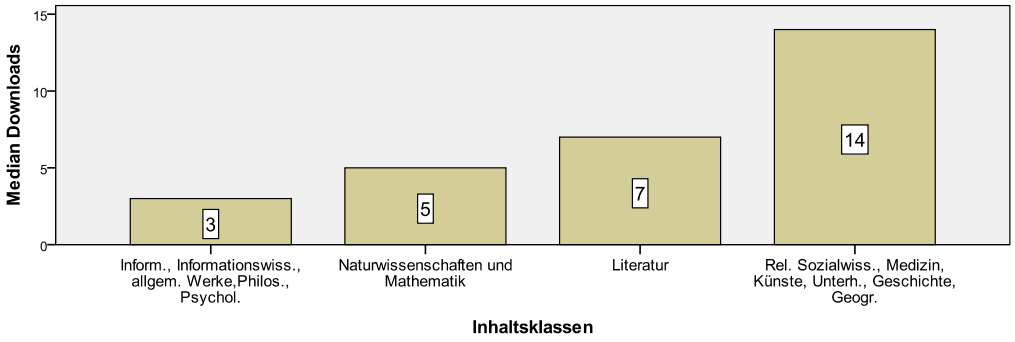


Abb. 69: Mediane der Downloads nach Inhaltsklasse, HeiDOK 12/2010

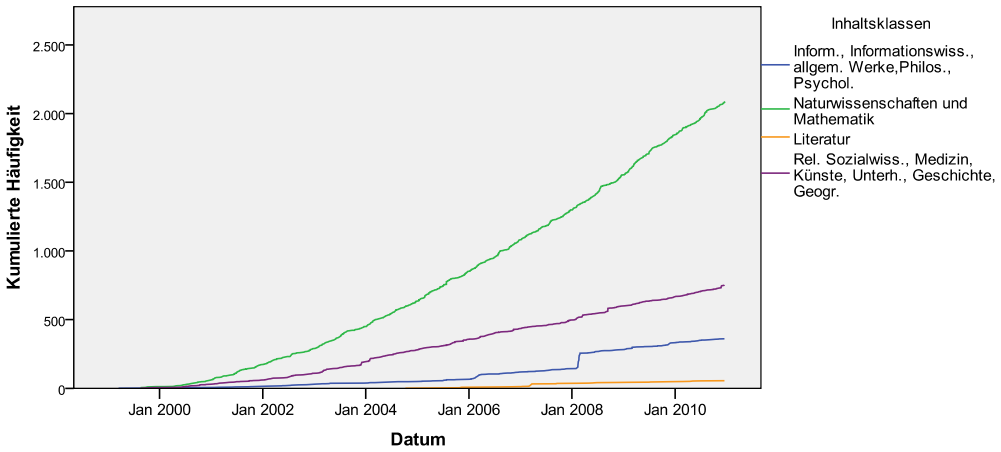


Abb. 70: Entwicklung des Bestandes nach Inhaltsklasse, HeiDOK 12/2010

5.3.2.3 Inhaltsklasse von Einzelpublikationen

Tab. 39: Häufigkeiten Inhaltsklasse von Einzelpublikationen, HeiDok 2010 Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	456	14,6	14,6	14,6
	1 Philosophie und Psychologie	660	21,2	21,2	35,8
	2 Religion	222	7,1	7,1	43,0
	3 Sozialwissenschaften	522	16,8	16,8	59,7
	4 Sprache	18	,6	,6	60,3
	5 Naturwissenschaften und Mathematik	780	25,0	25,0	85,4
	6 Technik, angewandte Wissenschaften	6	,2	,2	85,5
	61 Medizin	72	2,3	2,3	87,9
	7 Künste und Unterhaltung	90	2,9	2,9	90,8

	8 Literatur	174	5,6	5,6	96,3
	9 Geschichte und Geografie	114	3,7	3,7	100,0
	Gesamt	3114	100,0	100,0	

Die Hypothese H_0 wird abgelehnt. Um signifikante Unterschiede bei allen paarweisen Vergleichen zu erhalten, werden zusammengefasst: (0, 1, 5), (7, 8, 9), (2, 3).

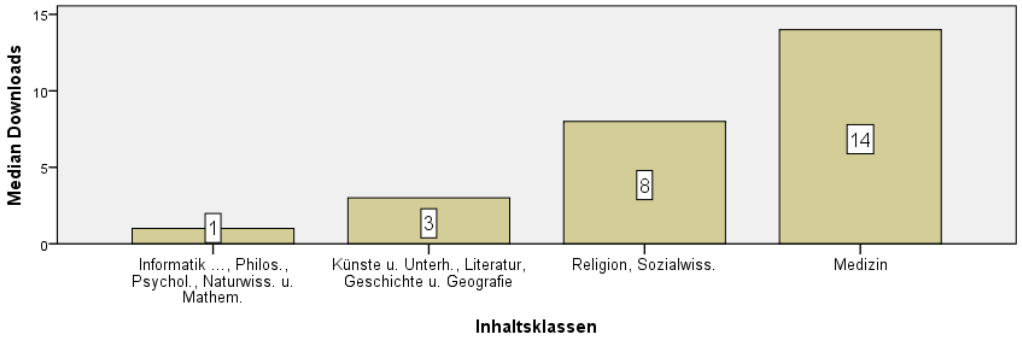


Abb. 71: Mediane der Downloads von Einzelpublikationen nach Inhaltsklasse, HeiDOK 12/2010

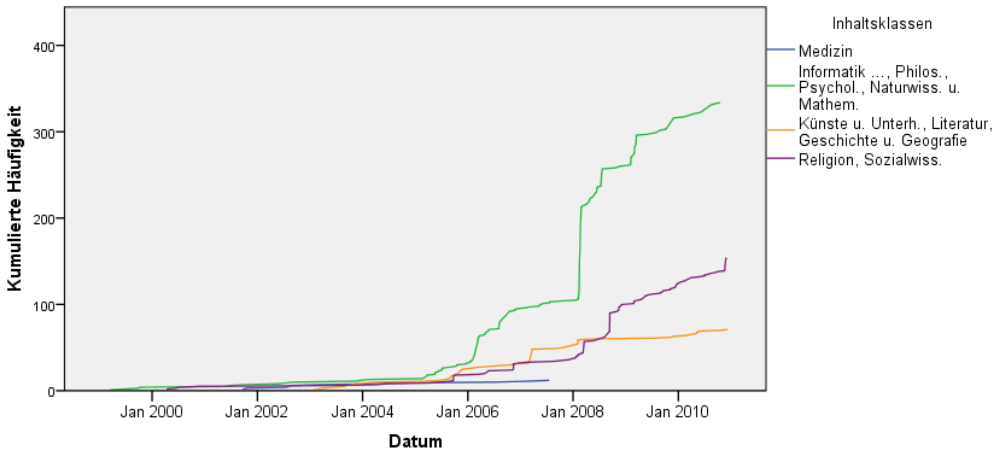


Abb. 72: Entwicklung des Bestandes an Einzelpublikationen nach Inhaltsklasse, HeiDOK 12/2010

5.3.2.4 Fakultät

Der Vergleich der Downloads nach Fakultäten ist wegen der Fallzahlen nur für Dissertationen sinnvoll. Es werden zwar für Abschlussarbeiten auch Ursprungsfakultäten angegeben, jedoch ist die Fallzahl nur in zwei Fakultäten ausreichend. Deshalb wird auf die Analyse der Abschlussarbeiten verzichtet.

Tab. 40: Häufigkeiten Fakultät für Dissertationen, HeiDok 2010 Downloadfile

F_Diss Fakultät Dissertation[a]					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 Fak. für Biowiss.	3654	28,9	28,9	28,9
	10 Neuphilologische Fakultät	144	1,1	1,1	30,1
	11 Philosophische Fakultät	510	4,0	4,0	34,1
	12 Theologische Fakultät	96	,8	,8	34,9
	13 Zentrale und Sonstige Einrichtungen	30	,2	,2	35,1
	2 Fak. f. Chemie und Geowiss.	2040	16,2	16,2	51,3
	3 Fak. f. Mathematik und Informatik	678	5,4	5,4	56,6
	4 Fak. f. Physik und Astronomie	4080	32,3	32,3	88,9
	5 Fak. f. Verhaltens- und Empirische Kulturwiss.	786	6,2	6,2	95,2
	6 Fak. f. Wirtschafts- und Sozialwiss.n	510	4,0	4,0	99,2
	7 Juristische Fakultät	6	,0	,0	99,2
	8 Medizinische Fakultät Heidelberg	72	,6	,6	99,8
	9 Medizinische Fakultät Mannheim	24	,2	,2	100,0
	Gesamt	12630	100,0	100,0	

Die Hypothese H_0 wird abgelehnt. Die Fakultäten 5, 6, 10 und 11 müssen zusammengefasst werden, um für alle signifikante paarweisen Vergleiche Signifikanz zu erhalten.

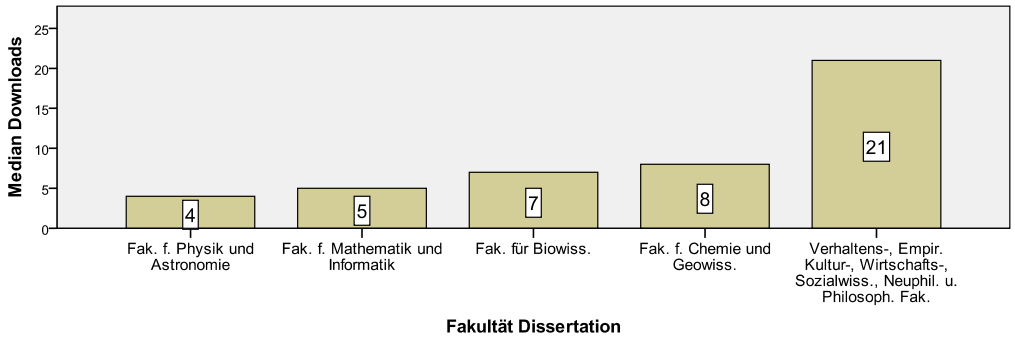


Abb. 73: Mediane der Downloads von Dissertationen nach Fakultät, HeiDOK 12/2010

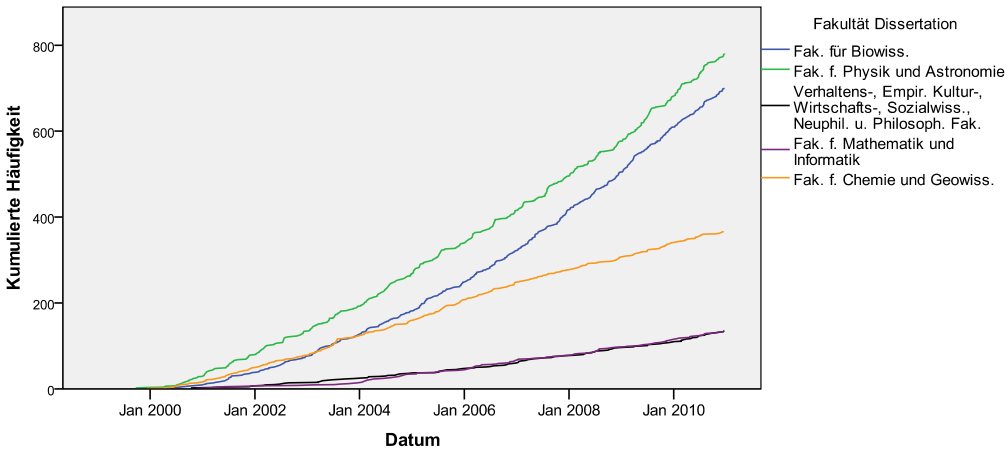


Abb. 74: Entwicklung des Bestandes von Dissertationen in Ergebnisgruppen nach Fakultät, HeiDOK 12/2010

5.3.3 Zusammenfassung

Formaler Publikationstyp

Die Publikationstypen mit den geringsten Anteilen, das sind Komplettpublikationen und Abschlussarbeiten, liegen bei den Downloads im oberen Bereich. Die Anzahl von beiden Typen wächst langsam. Dissertationen bilden mit 74 % den überwiegenden Anteil und ihre Downloads liegen im mittleren Bereich. Der Anteil von Einzelpublikationen ist mit 18 % geringer als bei den anderen IR und wächst langsam. Ihre Downloads liegen im unteren Bereich.

Inhaltsklasse

Die Gruppen der Downloads können grob in einen oberen und unteren Bereich eingeteilt werden. Der obere Bereich, der durch die Zusammenfassung von fünf Inhaltsklassen gebildet wird, ist in seinem Anteil an der Gesamtzahl von Publikationen niedrig und wächst langsam. Publikationen der Klasse „Naturwissenschaften und Mathematik“ bilden den größten Anteil und liegen bei den Downloads im unteren Bereich. Die Klasse „Literatur“ befindet sich im Übergang zum oberen Bereich der Downloads und ist anteilmäßig gering. Bei den Einzelpublikationen stagniert die Anzahl der Publikationen der Klasse „Medizin“, die die höchsten Downloads aufweist.

Fakultät

Die Downloads können wieder in einen oberen und unteren Bereich eingeteilt werden. Im unteren befinden sich Dissertationen aus Fakultäten des naturwissenschaftlichen und mathematischen Bereichs, die den überwiegenden Anteil an Dissertationen bilden. Im oberen Bereich der Downloads mussten vier Fakultäten zusammengefasst werden, die einen geringen Anteil an den Dissertationen einnehmen.

Fazit

Dissertationen sind der Anzahl nach die stärkste Gruppe und haben weniger Downloads als Abschlussarbeiten und Komplettpublikationen. Einzelpublikationen als zweitgrößte Kategorie haben die geringsten Downloads. Bei den Dissertationen bilden die aus den naturwissenschaftlichen und mathematischen Bereichen den größten Anteil, haben aber geringere Downloads als Dissertationen aus anderen Bereichen. Verglichen mit den anderen IR gibt es bei HeiDOK nicht den Effekt, dass die Anzahl von Publikationen in bestimmten Kategorien so stark steigt, dass die Anteile verändert werden. Die Fakultäten sind bei Dissertationen in vielen Fällen nicht die, die als Ursprungsfakultät angegeben sind und ausgewertet wurden, was dem Nutzer nicht auf dem ersten Blick klar ist.

Vergleich mit HeiDOK 7/2009

Nach wie vor unterscheiden sich die Downloads der vier Publikationstypen signifikant. Die Mediane haben sich nur geringfügig geändert und die Reihenfolge bleibt bestehen. Die Downloads für Abschlussarbeiten haben sich etwas erhöht, die von Dissertationen und Komplettpublikationen etwas verringert. Die Bildung von Gruppen von Fakultäten, deren Downloads sich signifikant unterscheiden, variiert. So konnten 2009 die Fakultäten für Wirtschafts- und Sozialwissenschaften als extra Gruppe identifiziert werden, deren Downloads niedriger lagen als die in den Fakultäten, mit denen sie 2010 zusammengefasst werden mussten. Die Tendenzen der Downloads für die Gruppen oder einzelne Fakultäten sind generell in beiden Analysen gleich.

5.4 Ergebnisse von SciDok (Saarbrücken)

5.4.1 Analyse der Metadaten

Analysiert werden die Metadaten von 1678 Publikationen.

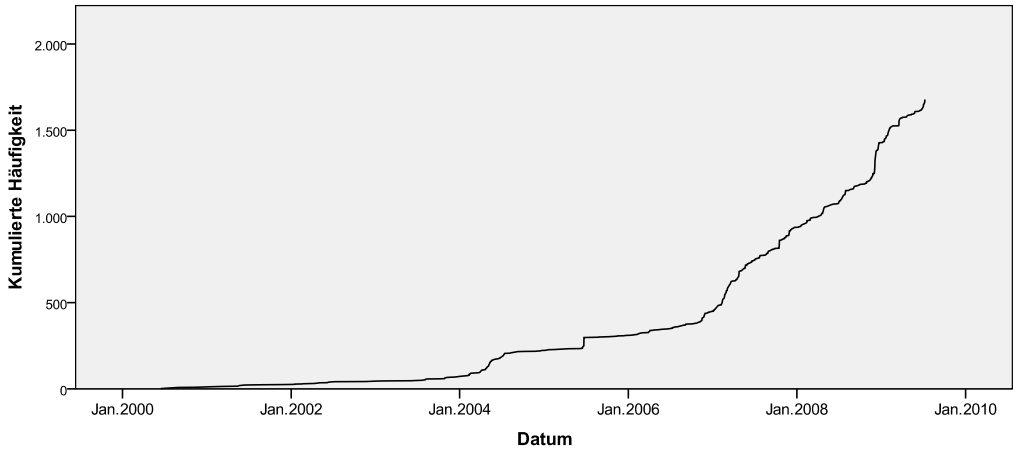


Abb. 75: Entwicklung des Bestandes an Publikationen, SciDok 6/2009

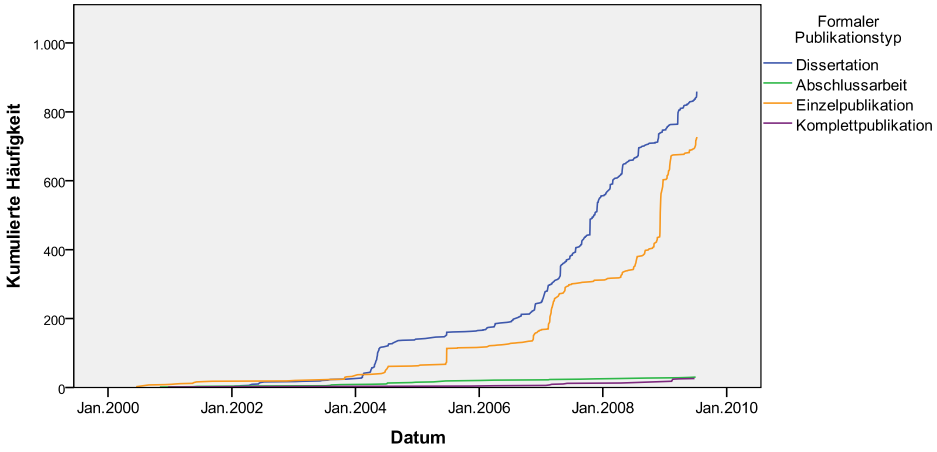


Abb. 76: Entwicklung des Bestandes an Publikationen nach Publikationstyp, SciDok 6/2009

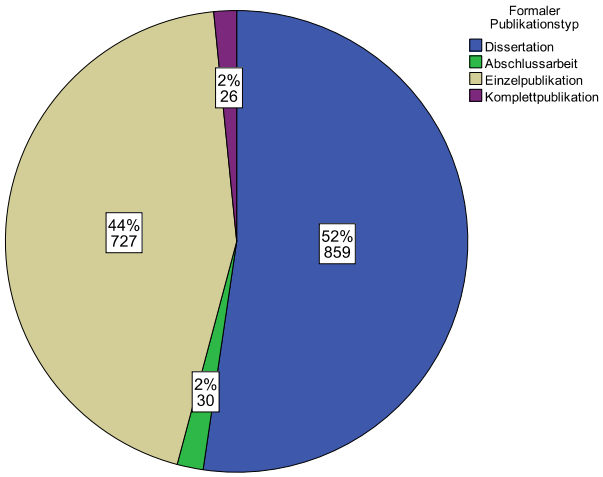


Abb. 77: Verteilung des Publikationstyps, SciDok 6/2009

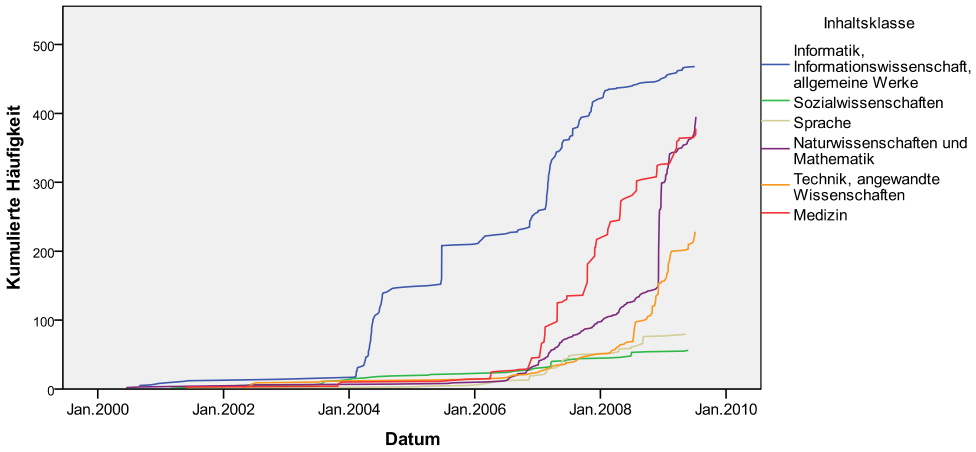


Abb. 78: Entwicklung des Bestandes an Publikationen nach Inhaltsklasse, SciDok 6/2009

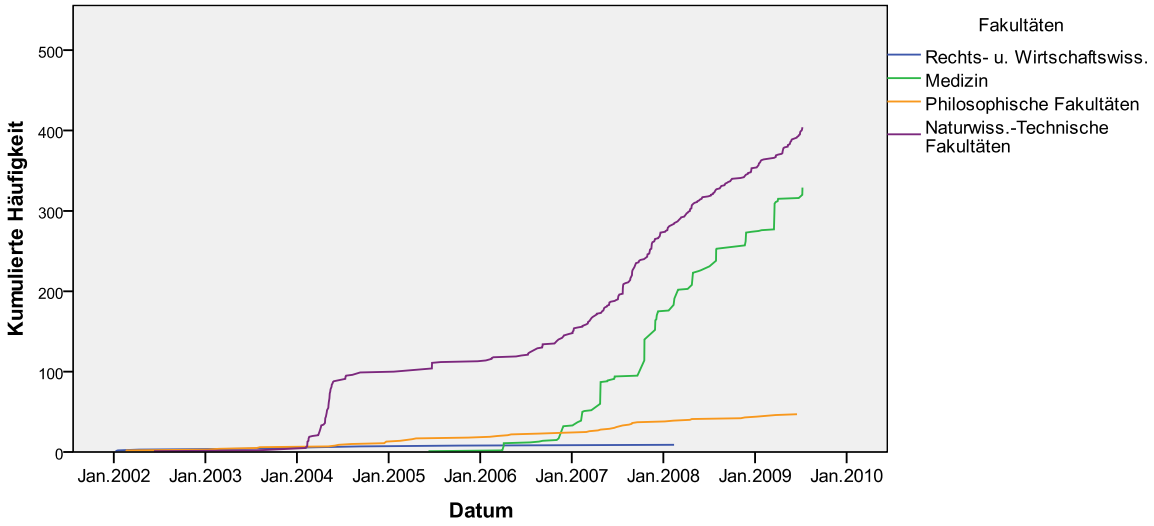


Abb. 79: Entwicklung des Bestandes von Dissertationen nach Fakultät, SciDok 6/2009

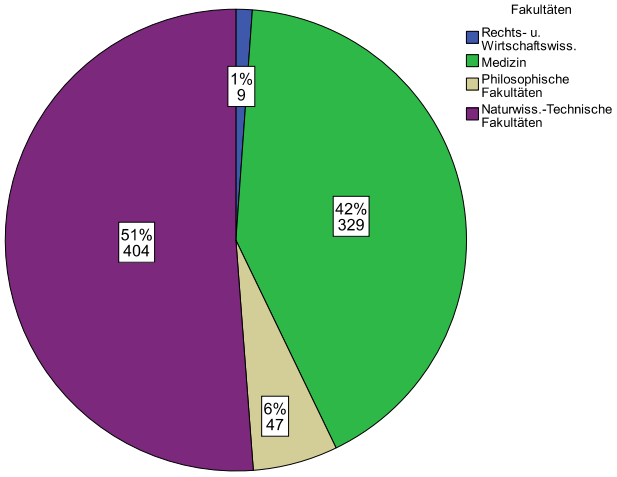


Abb. 80: Verteilung von Dissertationen nach Fakultät, SciDok 6/2009

5.4.2 Analyse der Nutzungsdaten

Analysiert werden die Downloads von 1076 Publikationen.

5.4.2.1 Formaler Publikationstyp

Tab. 41: Häufigkeiten Publikationstyp, SciDok Downloadfile

P_Typ Formaler Publikationstyp					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Dissertation	3960	61,3	63,2	63,2
	B Habilitation	6	,1	,1	63,3
	C Abschlussarbeit	156	2,4	2,5	65,7
	D Festschrift	6	,1	,1	65,8
	E Einzelpublikation	2064	32,0	32,9	98,8
	F Komplettpublikation	78	1,2	1,2	100,0
	Gesamt	6270	97,1	100,0	
Fehlend		186	2,9		
Gesamt		6456	100,0		

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Klassen sind signifikant. Die Kategorien „Dissertation“ und „Komplettpublikation“ unterscheiden sich nicht. Diese zusammenzufassen erscheint nicht sinnvoll, auch weil es nur 13 Komplettpublikationen gegenüber von 3960 Dissertationen gibt. Deshalb wird „Komplettpublikation“ in der Darstellung der Mediane weggelassen.

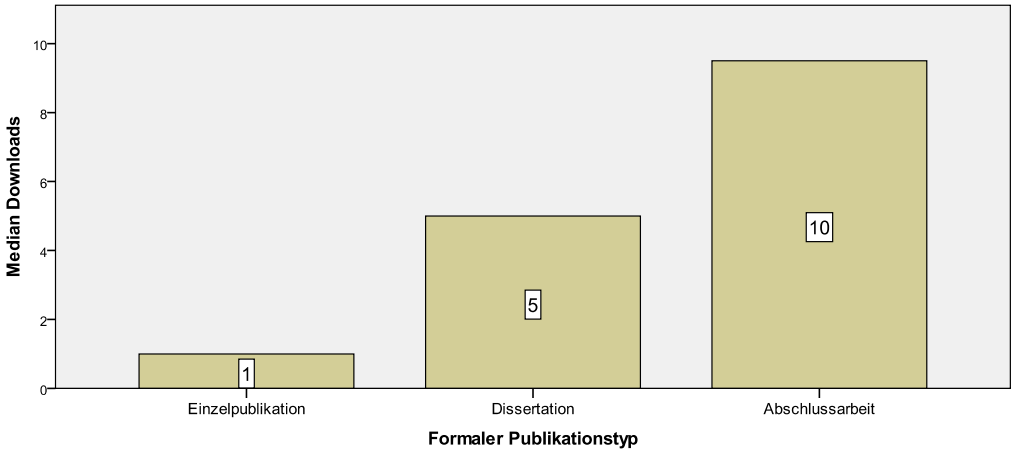


Abb. 81: Vergleich der Mediane nach Publikationstyp, SciDok 2009

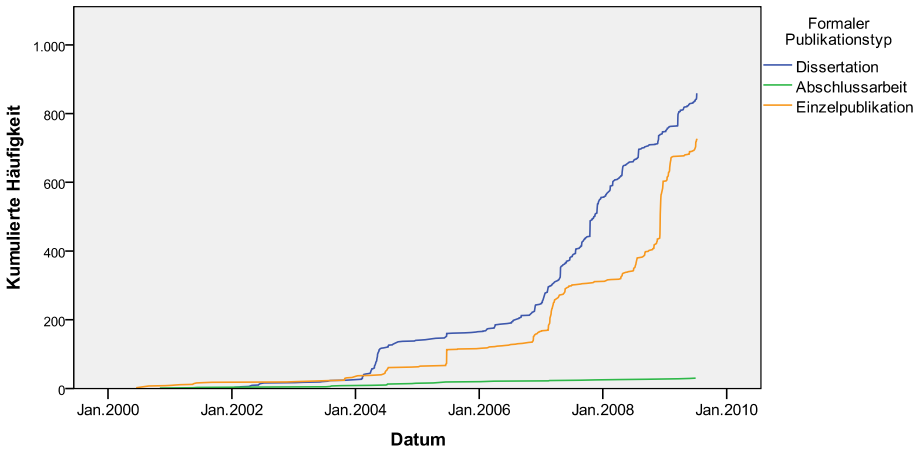


Abb. 82: Entwicklung des Bestandes in den Ergebnisgruppen nach Publikationstyp, SciDok 6/2009

5.4.2.2 Inhaltsklasse

Tab. 42: Häufigkeiten Inhaltsklasse, SciDok Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	2628	40,7	40,8	40,8
	1 Philosophie und Psychologie	138	2,1	2,1	43,0
	2 Religion	12	,2	,2	43,2
	3 Sozialwissenschaften	288	4,5	4,5	47,6
	4 Sprache	366	5,7	5,7	53,3
	5 Naturwissenschaften und Mathematik	756	11,7	11,7	65,1
	6 Technik, angewandte Wissenschaften	408	6,3	6,3	71,4
	61 Medizin	1656	25,7	25,7	97,1
	7 Künste und Unterhaltung	48	,7	,7	97,9
	8 Literatur	36	,6	,6	98,4
	9 Geschichte und Geografie	102	1,6	1,6	100,0
	Gesamt	6438	99,7	100,0	
Fehlend		18	,3		
Gesamt		6456	100,0		

Die Hypothese H_0 wird abgelehnt, aber nicht alle paarweisen Vergleiche der Kategorien sind signifikant. Alle paarweisen Vergleiche sind signifikant, wenn 1, 3, 5, 6 und 61 zusammengefasst werden.

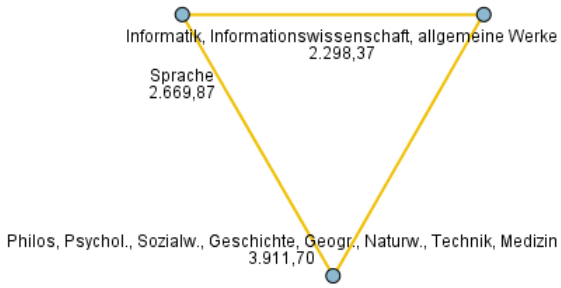


Abb. 83: Paarweise Vergleiche Inhaltsklasse, SciDok 6/2009

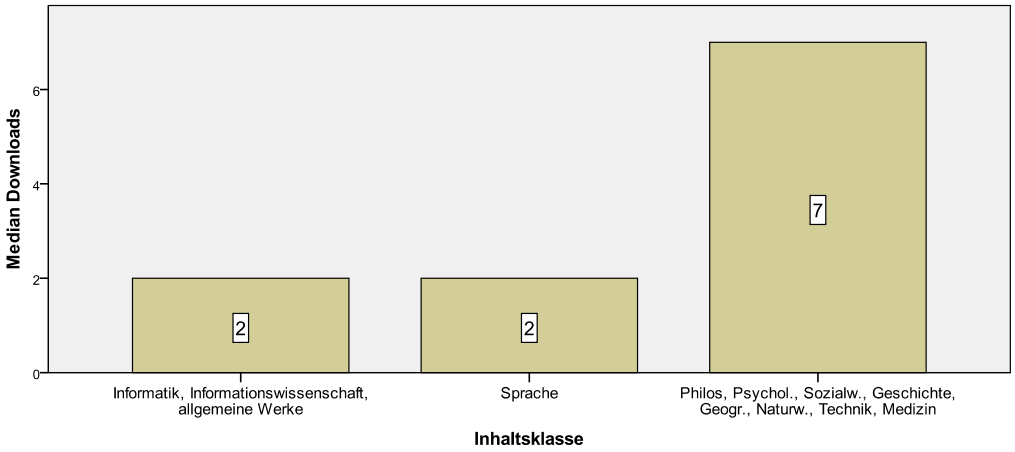


Abb. 84: Mediane der Downloads nach Inhaltsklasse, SciDok 6/2009

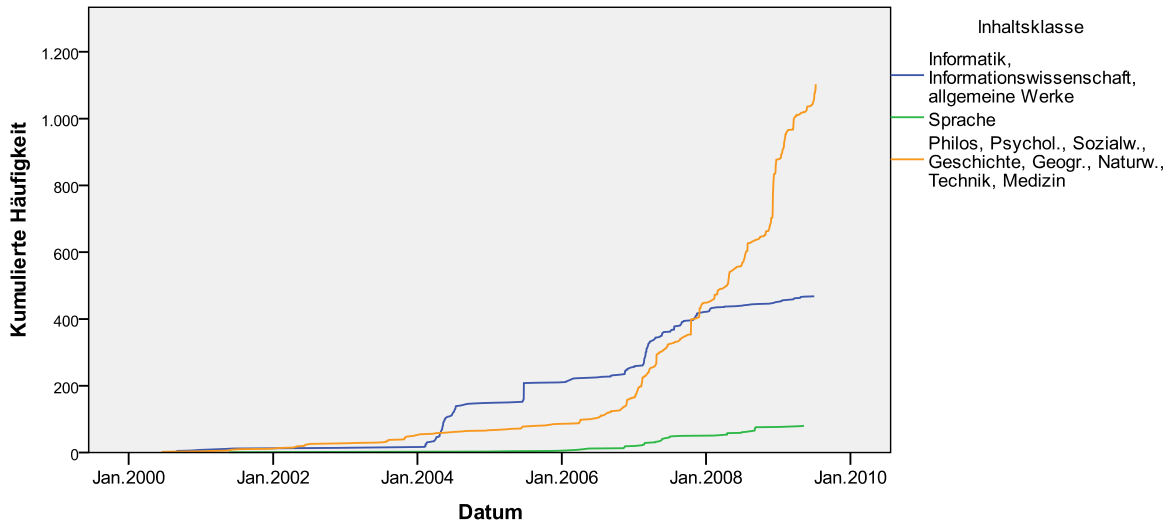


Abb. 85: Entwicklung des Bestandes in den Ergebnisgruppen nach Inhaltsklasse, SciDok 6/2009

5.4.2.3 Inhaltsklasse Einzelpublikationen

Tab. 43: Häufigkeiten Inhaltsklasse von Einzelpublikationen, SciDok Downloadfile

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	1278	61,9	62,5	62,5
	1 Philosophie und Psychologie	54	2,6	2,6	65,1
	3 Sozialwissenschaften	162	7,8	7,9	73,0
	4 Sprache	222	10,8	10,9	83,9
	5 Naturwissenschaften und Mathematik	36	1,7	1,8	85,6
	6 Technik, angewandte Wissenschaften	24	1,2	1,2	86,8
	61 Medizin	222	10,8	10,9	97,7
	8 Literatur	24	1,2	1,2	98,8
	9 Geschichte und Geografie	24	1,2	1,2	100,0
	Gesamt	2046	99,1	100,0	
Fehlend		18	,9		
Gesamt		2064	100,0		

Fasst man 0 und 4 zu einer neuen Kategorie und 3, 6 und 61 zu einer weiteren Kategorie zusammen, erhält man signifikante paarweise Vergleiche.

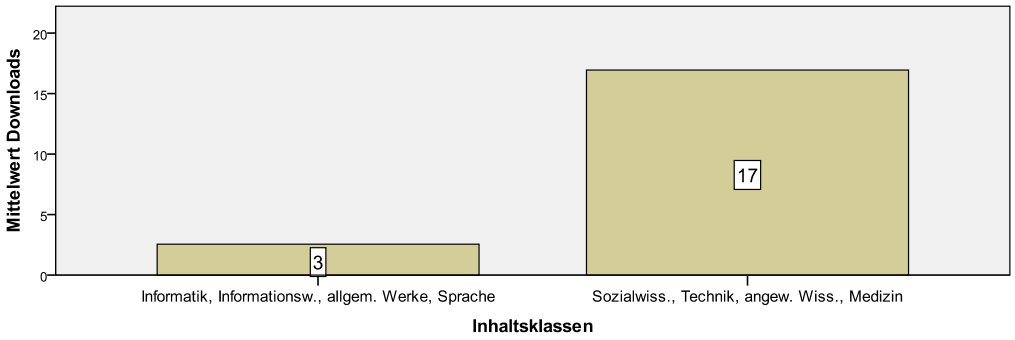


Abb. 86: Mediane der Downloads von Einzelpublikationen nach Inhaltsklasse, SciDok 6/2009

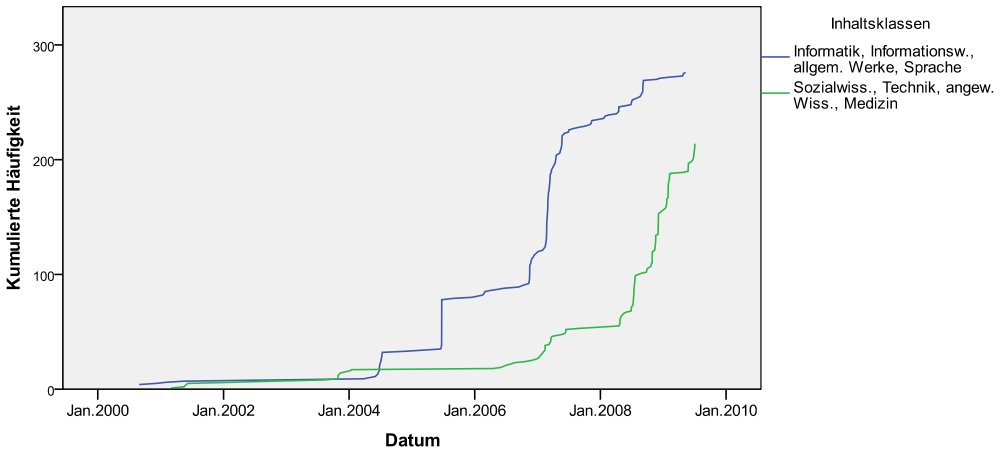


Abb. 87: Entwicklung des Bestandes an Einzelpublikationen in Ergebnisgruppen nach Inhaltsklasse, SciDok 6/2009

5.4.2.4 Fakultät

Tab. 44: Häufigkeiten Fakultät, SciDok Downloadfile

Fakultaet Fakultät					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 Fak. 1 - Rechts-und Wirtschaftswissenschaft	96	1,5	1,5	1,5
	2 Fak. 2 - Medizin	1620	25,1	25,5	27,0
	3 Fak 3 - Philosophische Fakultät I	42	,7	,7	27,6
	4 Fak 4 - Philosophische Fakultät II	294	4,6	4,6	32,3
	5 Fak 5 - Philosophische Fakultät III	1302	20,2	20,5	52,7
	6 Fak. 6 - Naturwissenschaftlich-Technische Fakultät I	1494	23,1	23,5	76,2
	7 Fak. 7 - Naturwissenschaftlich-Technische Fakultät II	60	,9	,9	77,2
	8 Fak. 8 - Naturwissenschaftlich-Technische Fakultät III	834	12,9	13,1	90,3
	9 Zentrale Einrichtungen	168	2,6	2,6	92,9
	10 Sonstige Einrichtungen	324	5,0	5,1	98,0
	11 MIssing	126	2,0	2,0	100,0
	Gesamt	6360	98,5	100,0	
Fehlend		96	1,5		
Gesamt		6456	100,0		

Wegen der geringen Fallzahlen in 3 und 7 und dem sachlogischen Zusammenhang werden die Philosophischen und Naturwissenschaftlichen Fakultäten jeweils zusammengefasst. Anders als bei der anderen IR gibt

es bei einer Mehrheit der Publikationen die Zuordnung zu einer Fakultät. Deshalb werden nicht nur Dissertationen, sondern auch Einzelpublikationen untersucht.

5.4.2.5 Fakultät von Dissertationen

Tab. 45: Häufigkeiten Fakultät von Dissertationen, SciDok 6/2009

Fak Fakultäten					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Rechts- u. Wirtschaftswiss.	54	1,4	1,5	1,5
	B Medizin	1356	34,2	38,0	39,6
	C Philosophische Fakultäten	246	6,2	6,9	46,5
	D Naturwiss.-Technische Fakultäten	1908	48,2	53,5	100,0
	Gesamt	3564	90,0	100,0	
Fehlend		396	10,0		
Gesamt		3960	100,0		

Die Hypothese H_0 wird abgelehnt und alle paarweisen Vergleiche der Fakultäten sind signifikant.

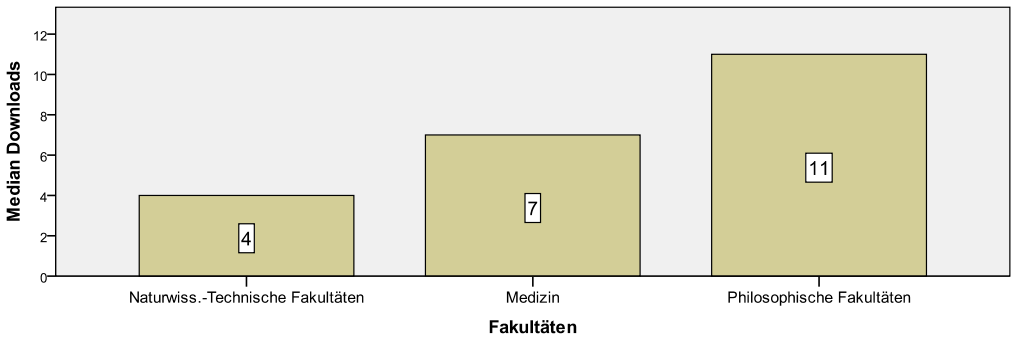


Abb. 88: Mediane der Downloads von Dissertationen nach Fakultäte, SciDok 6/2009

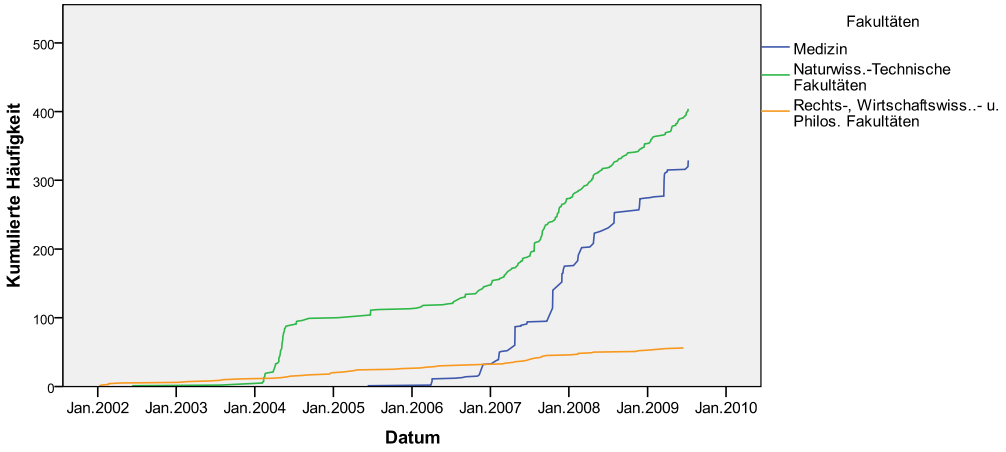


Abb. 89: Entwicklung des Bestandes an Dissertationen in Ergebnisgruppen nach Fakultät, SciDok 6/2009

5.4.2.6 Fakultät von Einzelpublikationen

Tab. 46: Häufigkeiten Fakultät von Einzelpublikationen, SciDok Downloadfile

Fak_E Fakultäten					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	A Rechts- u. Wirtschaftswiss.	36	1,7	2,0	2,0
	B Medizin	222	10,8	12,3	14,2
	C Philosophische Fakultäten	1128	54,7	62,3	76,5
	D Naturwiss.-Technische Fakultäten	426	20,6	23,5	100,0
	Gesamt	1812	87,8	100,0	
Fehlend		252	12,2		
Gesamt		2064	100,0		

Die Hypothese H_0 wird abgelehnt und alle paarweisen Vergleiche der Fakultäten sind signifikant.

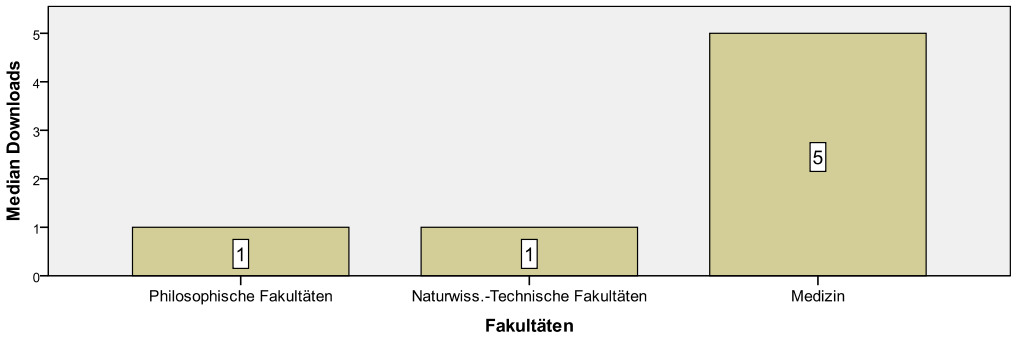


Abb. 90: Mediane der Downloads von Einzelpublikationen nach Fakultät, SciDok 6/2009

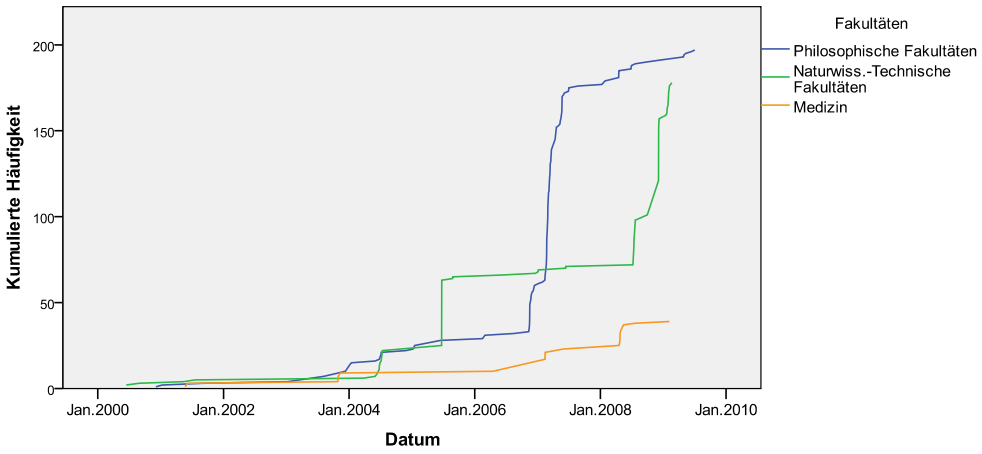


Abb. 91: Entwicklung des Bestandes an Einzelpublikationen in Ergebnisgruppen nach Fakultät, SciDok 6/2009

5.4.3 Zusammenfassung

Formaler Publikationstyp

Der Publikationstyp mit dem geringsten Anteil an der Gesamtzahl der Publikationen „Abschlussarbeit“ hat die meisten Downloads. Die Downloads der Dissertationen befinden sich im mittleren, die der Einzelpublikationen im unteren Bereich.

Inhaltsklasse

Aufgrund der geringen Fallzahlen konnten für die Kategorien „Sprache“ und „Informatik, Informationswissenschaften, allgemeine Werke“ und eine Zusammenfassung mehrerer anderer Kategorien (1, 3, 5, 6, 61) Downloads unterschieden werden. Die Anzahl der Publikationen in der Kategorie „Sprache“, die sich im unteren Bereich der Downloads befindet, wächst schnell und vergrößert ihren Anteil an der Gesamtzahl der Publikationen. Die Kategorie „Informatik, Informationswissenschaften, allgemeine Werke“ hat den größten Anteil an der Gesamtzahl der Publikationen, gehört aber bei den Downloads in den unteren Bereich. Zu den Kategorien der Publikationen, die sich im oberen Bereich der Downloads befinden, gehören mit „Medizin“ und „Naturwissenschaften und Mathematik“ Kategorien, die in der Anzahl gleichmäßig wachsen und ihren Anteil erhöhen.

Betrachtet man nur die Einzelpublikationen, können zwei Gruppen von Inhaltsklassen für die Downloads bestimmt werden. In der unteren Gruppe befinden sich wieder „Sprache“ und „Informatik, Informationswissenschaften, allgemeine Werke“, die schneller als die oberen Gruppe wachsen.

Fakultät

Die Zusammenfassung der Fakultäten erleichtern die Übersicht. Die Downloads der Dissertationen der Naturwissenschaftlich-Technischen Fakultät, deren Anteil an allen Dissertationen am größten ist, liegen im unteren, die Downloads der Medizinischen Fakultät im mittleren und die der Philosophischen Fakultäten,

deren Anteil am geringsten ist, im hohen Bereich. Da bei SciDok für alle Publikationen eine Fakultät oder Einrichtung angegeben wurde, können auch die Einzelpublikationen als zweite große Kategorie nach Publikationstyp untersucht werden. Die Publikationen der Philosophischen Fakultät haben den größten Anteil und befinden sich im unteren Downloadbereich. Ebenso im unteren Bereich befinden sich die Downloads der Naturwissenschaftlich-Technischen Fakultäten, deren Anzahl und Anteil steil ansteigen.

Fazit

Dissertationen haben den größten Anteil an den Publikationen und befinden sich bei den Downloads im mittleren Bereich. Innerhalb der Dissertationen haben die Naturwissenschaftlich-Technischen Fakultäten den größten Anteil und befinden sich bei den Downloads im unteren Bereich. Innerhalb der Einzelpublikationen ist der Anteil der Philosophischen Fakultäten am größten und die Downloads liegen im unteren Bereich. Es können keine generellen Aussagen darüber gemacht werden, ob die Gruppen mit geringeren Downloads zahlenmäßig stärker wachsen als die mit Downloads im oberen Bereich. Insgesamt ist es bei SciDok schwierig, Aussagen zu treffen, da die Unterschiede der Downloads weniger differenziert sind als bei den anderen IR. Das wird durch die Zusammenfassung von Kategorien deutlich. Eine Maßnahme, die Sichtbarkeit von Publikationen und damit des gesamten IR zu erhöhen, kann, wie bei edoc, sein, dem Nutzer bereits auf der Homepage des Browsings die Auswahl eines Fachgebietes oder einer Fakultät zu ermöglichen.

6 Diskussion

Die Analysemethode NoRA für Institutional Repositories (IR) wurde anhand des Datenmaterials von vier universitären IR entwickelt, erprobt und mit den im vorangegangenen Kapitel dargestellten Ergebnissen auf diese angewendet. Im ersten Teil dieses Kapitels wird die Leistungsfähigkeit von NoRA im Hinblick auf die Zielstellung der vorgelegten Arbeit bewertet. Im zweiten Teil wird gezeigt, dass durch die Anwendung von NoRA nicht nur das einzelne IR charakterisiert werden kann, sondern dass darüberhinaus aus den vier Fallbeispielen Zusammenhänge zwischen der Struktur und Nutzung des IR abgeleitet werden können, die so bisher nicht bekannt waren. Aus den Erkenntnissen ergeben sich Einsichten in das Verhalten von Nutzern, die Konsequenzen für den Betrieb von IR und die Interpretation von deren Nutzungsdaten haben und die weitere Anwendungsfelder von NoRA aufzeigen.

6.1 Eine Bewertung von NoRA

Ziel der Arbeit war es, eine Analysemethode für Nutzungsdaten von IR zu entwickeln, welche es ermöglicht, die Sichtbarkeit von Publikationsgruppen im Internet verlässlich zu charakterisieren. Mit nichtparametrischen Verfahren der Mathematischen Statistik, der Hauptkomponente der hierzu entwickelten Methode NoRA, wurden Mittel gefunden, wie Zugriffszahlen, die bekanntermaßen stark schwanken und unkontrollierten Einflüssen unterliegen, systematisch ausgewertet werden können. NoRA identifiziert Publikationsgruppen, welche sich in ihrer Nutzung signifikant unterscheiden. Die für den Vergleich der Nutzung von Publikationsgruppen geeignete Kenngröße ist der Median der Downloads. Zusammen mit der Analyse des Bestandes an Publikationen anhand der Metadaten ergeben die Mediane ein Bild der Struktur des IR und dessen Sichtbarkeit im Internet. Diese Informationen ermöglichen es den Betreibern von IR gezielte Maßnahmen zur qualitativen Verbesserung des Repository zu ergreifen.

Die Publikationsgruppen wurden auf der Grundlage von inhaltlichen und formalen Merkmalen definiert, die in den Metadaten DINI-zertifizierter Repositories generell enthalten sind. Durch geeignete Zusammenfassungen wurden Klassifikationen geschaffen, welche eine einheitlich gestaltete Analyse von IR ermöglichen. Da formale Publikationstypen mit verschiedenen Bezeichnungen auftreten und selbst bei gleicher Bezeichnung unterschiedlich interpretiert werden, konnten zum Teil nur stark vereinfachte Kategorien gefunden werden. Das hat den Vorteil, dass die Kategorien „Einzelpublikation“ und „Komplettpublikation“ vermutlich für alle IR anwendbar sind, aber auch den Nachteil, dass individuelle Details nicht aufgedeckt werden. Dieser zuletzt genannte Mangel kann allerdings dadurch behoben werden, dass nicht die in NoRA vorgeschlagenen verallgemeinerten Klassifikationen, sondern die in den Metadaten original vorhandenen in die Analyse eingehen. Das Beispiel ehsStu, bei dem zusätzlich die Kategorie „Studienarbeit“ berücksichtigt wurde, die bei den übrigen drei IR nicht vertreten war, zeigt die Flexibilität von NoRA. Ebenso flexibel ist die Anwendung auf Publikationsgruppen, die durch Fakultäten gebildet werden, die bei allen IR unterschiedliche Kategorien enthalten. Gerade hier wird deutlich, wie NoRA auf spezielle Anforderungen angepasst werden kann.

Obwohl NoRA auf beliebige Klassifikationen angewendet werden kann, ist es wünschenswert - auch im Hinblick auf einen späteren Vergleich von IR - mit einheitlichen Kategorien zu arbeiten. Im Fall einer formalen Klassifikation wurde mit dem von DINI vorgeschlagenem gemeinsamen Vokabular für Publikationstypen (DINI 2010b) dafür prinzipiell die Voraussetzung geschaffen. Die konsequente Umsetzung, welche die semantisch korrekte Verwendung der Begriffe einschließt, würde zu einer qualitativen Verbesserung der Analyse führen.

Da eine inhaltliche Klassifikation bei allen vier analysierten IR durch die Vergabe von DDC-Werten vorliegt, wird auf die 10 Hauptklassen der DDC zurückgegriffen. Lediglich „Medizin“ wird aus der Klasse „Technik, angewandte Wissenschaften“ isoliert und bildet eine eigene zusätzliche Klasse, da in diesem Fall eindeutig eine Sonderstellung durch die hohe Anzahl der Publikationen vorliegt. Eine Klassifikation nach Fachgebieten, wie es z. B. bei OpenDOAR vorgesehen ist, wäre wünschenswert. Es wird deutlich, dass eine inhaltliche Klassifikation der Publikationen nach DDC für die Beschreibung eines IR nur bedingt geeignet ist, wenn sie mit dem Ziel erfolgt, herauszufinden, wo verstärkt Publikationen akquiriert werden sollten. Mit der Erfassung der Fakultät für alle Publikationen wäre es besser möglich, die Ergebnisse von NoRA auf Autoren und Herausgeber zu beziehen.

Für die Datenauswahl werden in NoRA feste Regeln vorgegeben, die leicht zu befolgen sind und keine aufwendigen Datenmanipulationen erfordern. Dadurch wird die Analyse erheblich vereinfacht. Das gilt sowohl bei der Auswahl der zugelassenen Kategorien für die Bildung signifikant unterschiedlicher Publikationsgruppen nach Downloads als auch bei der Auswahl der zu analysierenden Nutzungsdaten. Die beste Voraussetzung für die Anwendung von NoRA sind monatliche Downloads. Die Auswertung der Daten von ehsStu zeigt jedoch, dass eine erfolgreiche Analyse möglich ist, wenn anders aggregierte Nutzungsdaten vorliegen.

Wegen der charakteristischen Eigenschaften der Nutzungsdaten ist es unumgänglich, Verfahren der Mathematischen Statistik für ihre Auswertung anzuwenden, da nur so eine begründbare Bildung von Publikationsgruppen und die Angabe von Vergleichswerten möglich sind. Da mit SPSS ein Statistikprogramm gefunden wurde, welches mit geringem Einarbeitungsaufwand das Datenmanagement, die Durchführung der statistischen Tests zur Bildung von Publikationsgruppen und die grafische Darstellung der Ergebnisse bietet, kann erwartet werden, dass NoRA auf Akzeptanz bei den Betreibern von IR trifft, auch wenn diese nicht mit der statistischen Datenanalyse vertraut sind. Die Auswertung des Kruskal-Wallis-Tests und die anschließend eventuell notwendige Zusammenfassung von Kategorien der Gruppenvariablen erfordert dann besondere Sorgfalt, wenn mehr als fünf Kategorien vorliegen und nicht alle paarweisen Vergleiche signifikant ausfallen. Hierbei zahlt sich ein inhaltliches Verständnis des in NoRA verwendeten Prinzips der statistischen Datenanalyse aus. Dieses Verständnis kann durch Befolgung der einzelnen Schritte mit eigenen Daten und am Beispiel der analysierten vier IR erworben werden.

6.2 Ergebnisse von NoRA

Die im vorangegangenen Kapitel vorgestellten Ergebnisse unterscheiden sich im Detail und bilden spezifische Eigenschaften der IR ab. Sie zeigen aber auch Gemeinsamkeiten, welche auf Probleme hinweisen, die nicht nur bei diesen vier IR, sondern genereller Art sein dürften und daher mehr Beachtung finden sollten, als es in der gegenwärtigen Praxis der Fall ist. Gleichzeitig werfen sie ein Licht darauf, was Nutzungsdaten aussagen können und wie sie nicht zu interpretieren sind.

6.2.1 Akzeptanz als Publikationsmedium

Anhand der Entwicklung des Publikationsbestandes lässt sich abbilden, wie sich die Akzeptanz des IR als Publikationsmedium bei Autoren und Herausgebern verändert. Die an vier universitären IR dazu durchgeführten Untersuchungen können nicht als repräsentativ gelten, weisen aber auf allgemeine Veränderungen hin.

Alle vier vorgestellten IR sind aus Dokumentenservern von Universitäten hervorgegangen, deren Aufgabe es ist, wissenschaftliche Publikationen zu archivieren und frei zur Verfügung zu stellen. Dabei handelte es sich am Anfang vor allem um typische Hochschulschriften wie Dissertationen und Habilitationen[133]. Ausgelöst wurde diese Entwicklung Ende der 1990er Jahre durch die Möglichkeit, der Publikationspflicht für Dissertationen durch eine elektronische Veröffentlichung nachzukommen. Dissertationen bilden seither einen wichtigen und großen Anteil an den Publikationen der IR. Ihre Anzahl wächst nach wie vor gleichmäßig. Im Laufe der Zeit veränderte sich das Spektrum der Publikationen und Aufgaben der IR und die Dokumentenserver entwickeln sich zunehmend zu Serviceeinrichtungen der Universitäten. Bei zwei der vier IR übersteigt der Anteil an anderen Publikationen bereits den der Dissertationen. Unter Autoren von Dissertationen ist und war die Akzeptanz eines IR als Publikationsmedium aus verschiedenen Gründen wie geringe Kosten, gute Verbreitungsmöglichkeit und Einfachheit des Verfahrens von Anfang an vorhanden. Das zeigt sich an der kontinuierlich wachsenden Anzahl von Dissertationen. Unter Autoren und Herausgebern von anderen Publikationen setzt sich die Überzeugung vom Sinn und den Vorteilen von OA langsamer durch. Das gilt sowohl für die Veröffentlichung von Pre- und Postprints einzelner Artikel und Beiträge als auch für die Veröffentlichung ganzer Zeitschriften und Serien. In letzter Zeit werden jedoch die Möglichkeiten der IR mehr und mehr für eine elektronische Zweitpublikation neben der Papierpublikation genutzt. Ein weiteres Zeichen für die steigende Akzeptanz von IR ist die wachsende Anzahl von Zeitschriften, die von der Papierform auf die ausschließlich elektronische Publikation transformiert oder von Beginn an als Online-Publikation konzipiert werden. Damit hält neben der „Green Road to Open Access“ auch das Modell der „Golden Road“ Einzug auf IR. Widergespiegelt wird diese Entwicklung im Anstieg von Einzelpublikationen, der sprunghaft verläuft,

[133] Aus den Daten von ehsStu geht nur die Entwicklung des Datenbestandes ab 14.6.1999 hervor. Es existieren 454 Datensätze mit diesem Datum.

wenn eine Zeitschrift erstmalig online publiziert wird. Um diese positive Entwicklung zu fördern, muss eine intensive Betreuung von Herausgebern stattfinden und auf individuelle Anforderungen eingegangen werden.

Für die Analyse wurden aus Gründen der Vereinheitlichung der bei den vier IR vorhandenen formalen Publikationstypen alle unselbständigen Publikationen und alle Typen von Publikationen wie Aufsätze, Berichte usw., die nicht Bestandteil einer selbständigen Publikation sein müssen, unter der Kategorie „Einzelpublikation“ zusammengefasst. Selbstständigen Publikationen, für die es einen kompletten Volltext aller Bestandteile gibt, wurden in die Kategorie „Komplettpublikation“ eingereiht (siehe Abschnitt 3.3.1). In allen vier Fällen bilden Einzelpublikationen und Dissertationen den größten Anteil am jeweiligen Bestand. Komplettpublikationen und andere Hochschulschriften als Dissertationen sind bei den vier IR dagegen unterschiedlich stark vertreten. Es ist zu vermuten, dass die Anzahl der Dissertationen weiterhin kontinuierlich steigen wird. Ein stärkerer Anstieg wird auch durch gezielte Maßnahmen, wie intensivierte Autorenbetreuung und Akquisition von Dissertationen aus Sachgebieten, die bisher unterrepräsentiert sind, nicht erreichbar sein, da die Gesamtzahl der jährlichen Schriften begrenzt ist. Bei Einzelpublikationen besitzen vor allem Zeitschriften großes Entwicklungspotential, da sie durch quantitative Veränderungen in der Anzahl von Publikationen zur qualitativen Verbesserung der IR maßgeblich beitragen.

6.2.2 Die Nutzung der Publikationen

Die Analyse der Downloads der IR hat gezeigt, dass sich die Nutzung der beiden großen Gruppen „Dissertation“ und „Einzelpublikation“ stark unterscheidet. Befinden sich Dissertationen immer im oberen Bereich der Downloads, so sind Einzelpublikationen regelmäßig im unteren Bereich zu finden. Geht man davon aus, dass alle vier IR ein DINI-Zertifikat besitzen und die formalen Bedingungen für die Herstellung der Sichtbarkeit wie die Existenz einer OAI-Schnittstelle für alle Publikationen gleichermaßen erfüllt sind, ist die Sichtbarkeit im Internet für alle Publikationstypen gleichermaßen gegeben. Die generelle Sichtbarkeit des Gesamtangebotes in der Institution ist ebenfalls sichergestellt. Es muss also andere Ursachen geben, warum bestimmte Publikationstypen mehr genutzt werden als andere.

Wie in Kapitel 1 dargelegt, geht das Konzept des Download Impact von Publikationen von der Annahme aus, dass durch die Nutzung ein Interesse des Rezipienten an der Publikation bekundet wird. Dabei wird das Konzept des Citation Impact, bei welchem davon ausgegangen wird, dass die Zitation eine Art von Belohnung darstellt und der Citation Impact als ein Maß für die Wirkung einer Publikation in Fachkreisen dient, auf die Nutzung von elektronischen Publikationen übertragen. Die Ergebnisse der Analyse der vier IR legen nahe, dass es außer dem Interesse am Inhalt noch andere Auslöser für den Download einer Publikation gibt.

Im Folgenden wird versucht zu klären, worin die Ursachen für die Unterschiede in den Downloads der Publikationen der vier IR liegen können. Die genauere Untersuchung, ob und wie stark das Nutzungsverhalten von den einzelnen Faktoren bestimmt wird, kann interessante Einsichten in das Verhalten von Nutzern geben, erfordert jedoch umfangreiche Tests und ist nicht Gegenstand der Arbeit. Für eine solche Untersuchung bietet sich die Verwendung von NoRA auch an.

6.2.2.1 Das Interesse an den Autoren

Vergleicht man die Downloads der Publikationstypen unter der Annahme, dass Nutzung Interesse am Inhalt ausdrückt, so besteht am Inhalt von Qualifikationsarbeiten wie Dissertationen, Habilitationen und Abschlussarbeiten bei allen vier IR größeres Interesse als an dem von Einzelpublikationen. Bezieht man den Publikationstyp „Festschrift“ mit ein, der bei edoc unter „Öffentliche Vorlesungen“ zu finden ist und die höchsten Downloads aufweist, wächst der Unterschied der Downloads zu den Einzelpublikationen noch einmal beträchtlich. Wie man in der folgenden Tabelle sieht, ist bei den Festschriften der Anteil aus der Inhaltsklasse „Sozialwissenschaften“ bei den Festschriften überproportional hoch. Die Klasse „Sozialwissenschaften“ hat, vergleicht man alle Publikationen von edoc, aber gerade den niedrigsten Median der Downloads.

Tab. 47: Häufigkeiten Inhaltsklasse von Festschriften, edoc 3/2010

I_Klasse Inhaltsklasse					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0 Informatik, Informationswissenschaft, allgemeine Werke	18	10,3	11,5	11,5
	1 Philosophie und Psychologie	10	5,7	6,4	17,9
	2 Religion	6	3,4	3,8	21,8
	3 Sozialwissenschaften	81	46,3	51,9	73,7
	4 Sprache	5	2,9	3,2	76,9
	5 Naturwissenschaften und Mathematik	2	1,1	1,3	78,2
	6 Technik, angewandte Wissenschaften	1	,6	,6	78,8
	8 Literatur	8	4,6	5,1	84,0
	9 Geschichte und Geografie	25	14,3	16,0	100,0
	Gesamt	156	89,1	100,0	
Fehlend		19	10,9		
Gesamt		175	100,0		

Bei Qualifikationsarbeiten kommt der überwiegende Teil aus naturwissenschaftlich-mathematischen und medizinischen Gebieten, welche in Hinsicht auf die Downloads aller Publikationen, nur wenig höher liegen als die Sozialwissenschaften.

Wie kann dieser Widerspruch, dass bei bestimmten Publikationstypen die Nutzung derartig vom allgemeinen Trend abweicht, erklärt werden? Sowohl bei Qualifikationsarbeiten als auch bei dem Typ „Festschrift“ handelt es sich um Publikationen, die einem einzigen Autor zugeordnet werden. Die Vermutung liegt nahe, dass bei den Rezipienten weniger Interesse am Inhalt der Publikation, sondern eher an der Publikation bestimmter Autoren besteht. Das würde die Annahme, auf dem der Download Impact beruht, in Frage stellen.

6.2.2.2 Der Einfluss der Position

Eine weitere Auffälligkeit ergibt sich bei der Betrachtung der Downloads von Abschlussarbeiten. Bei edoc unterscheiden sich die Downloads von Abschlussarbeiten nicht signifikant von denen von Dissertationen und Habilitationen und können daher nicht in die folgende Betrachtung einbezogen werden. Bei HeiDok und SciDok haben Abschlussarbeiten bedeutend höhere Downloads als Dissertationen. Warum sollten Abschlussarbeiten von größerem Interesse sein als Dissertationen? Auf der Website zur Auswahl von Dokumenttypen von HeiDok stehen Abschlussarbeiten an erster Stelle, bei SciDok zwar nicht an erster Stelle, aber vor Dissertationen. In der Tabelle werden die Positionen von „Abschlussarbeit“ und „Dissertation“, die Größen der Gruppen und die Mediane der Downloads verglichen. Bei HeiDok und SciDok ist das Verhältnis der Anzahl etwa gleich. Ebenso steht bei beiden IR „Abschlussarbeit“ auf der Website vor „Dissertation“. Der Median von „Abschlussarbeit“ ist jeweils etwa doppelt so hoch wie der von „Dissertation“.

Tab. 48: Übersicht über Position von Abschlussarbeiten und Dissertationen im Browsing

		ehsStu 8/2009	**HeiDOK 12/2010**	**SciDok 6/2009**
Position im Browsing	Abschlussarbeiten	3 und 5*	1	3 und 5*
	Dissertationen	6	6	6
Anzahl	Abschlussarbeiten	629	128	33
	Dissertationen	2013	2605	1315
Median Downloads	Abschlussarbeiten	123	11	10
	Dissertationen	174	6	5
Verhältnis der Mediane		1:1,4	1:1,8	1:2
Verhältnis der Anzahl rund		1:3	1:20	1:40

*Bachelor mit Position 3 und Diplomarbeit, Magisterarbeit mit Position 5

Es ist evident, dass die Position in einer Liste und damit in einer Rangfolge Einfluss auf die Nutzung (siehe auch Abschnitt 2.3.4 Sichtbarkeit und Ranking im Internet) ausübt. Das gilt auch für Zitationen von elektronischen Publikationen, denen die Nutzung vorausgeht. So untersuchten Haque und Ginsparg den Einfluss der Position von Artikeln in der Webankündigung von arXiv.org auf die Downloads und spätere Zitation (Haque und Ginsparg 2009) und stellten fest, dass die Mediane der Downloads mit der Position abnehmen. Dabei ist der Unterschied zwischen der 1. und 2. Position am größten. Die Sichtbarkeit, die durch eine zufällig entstandene Position in der Reihenfolge erzeugt wird, ist entscheidend für die Downloads. Gleiches gilt auch für die Downloads von Publikationen von IR. Hier wird unbeabsichtigt ein Ranking vorgenommen, welches sich auf die Downloads auswirkt. Die Position im Browsing, durch die alphabetische Sortierung der Dokumentarten bestimmt, beeinflusst die Nutzung. Dieser Effekt wurde schon bei früheren Analysen von Dissertationen bei edoc festgestellt. Durch die alphabetische Sortierung der Dissertationen nach dem Namen des Autors auf

der Website wird ein Ranking produziert, welches für Autoren, deren Nachname mit A beginnt, einen Sichtbarkeitsvorteil erzeugt und zu signifikant höheren Downloads führt, was die folgende Abbildung zeigt.

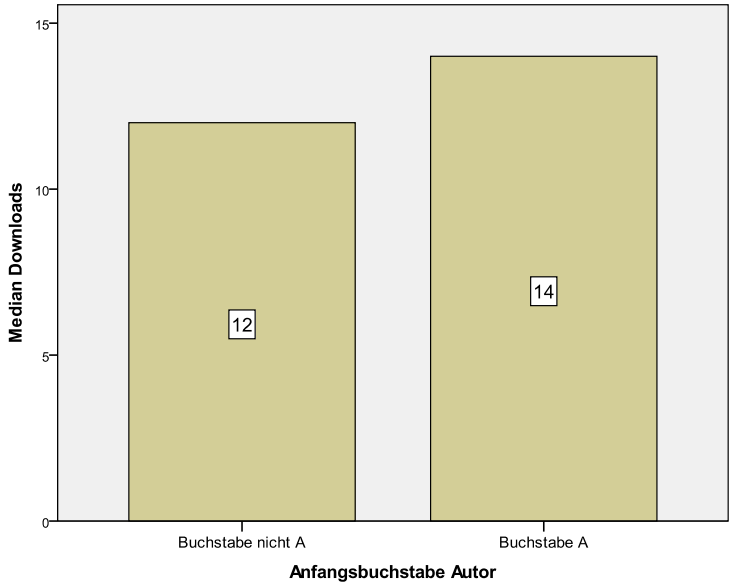


Abb. 92: Downloads von Dissertationen nach den Anfangsbuchstaben der Autoren, edoc 2008

6.2.2.3 Die Anzahl von Publikationen

Ein weiterer Faktor, der die Zahl der Downloads beeinflusst, ist die Größe der Gruppe, die der Rezipient beim Browsing in einem Fenster vorfindet. Bei ehsStu findet man im Browsing Abschlussarbeiten auch vor Dissertationen, aber Abschlussarbeiten haben hier weniger Downloads. Vergleicht man die Anzahl der Publikationen in den beiden Gruppen, unterscheidet sich das Größenverhältnis von 1:3 bei ehsStu beträchtlich von den Größenverhältnissen von 1:20 bzw. 1:40 von HeiDok und SciDoc (siehe Tabelle 48). Die Mediane von Abschlussarbeiten sind bei SciDok und HeiDok rund doppelt so hoch wie die Mediane der Dissertationen, während sich bei HeiDok die Mediane weit weniger stark unterscheiden. In der folgenden Grafik werden die Downloads von zwei Fakultäten mit stark unterschiedlicher Anzahl von Publikationen verglichen.

Abb. 93: Mediane der Downloads von zwei Fakultäten, SciDok 2009

Fakultät 6 mit 249 Publikationen steht im Browsing vor Fakultät 7 mit 10 Publikationen. Das Beispiel SciDok verdeutlicht, wie sich die Downloads von aufeinanderfolgenden Gruppen unterscheiden können, wenn die Gruppen in der Größe deutlich voneinander abweichen. Der Einfluss der Position wird durch den Einfluss der Größe der Gruppe überdeckt. Ein weiteres Beispiel, wie die Größe der Gruppe den Einfluss der Position aufhebt, ist der schon erwähnte Unterschied der Downloads von Dissertationen (Median = 12, Anzahl = 1600) und Festschriften (Median = 21, Anzahl = 169) bei edoc. Geht man davon aus, dass nur die Position im Browsing Einfluss hat, müssten die Dissertationen höhere Downloads aufweisen. Hier wiegt jedoch offenbar der starke Unterschied in der Gruppengröße den Einfluss der Position auf die Downloads auf. Dafür kann eine einfache Erklärung gefunden werden. Hat sich der Nutzer entschlossen, ein nächstes Fenster zu öffnen und auf eine konkrete Publikation zuzugreifen, ist die Wahrscheinlichkeit für eine einzelne Publikation, vom Nutzer gewählt zu werden, um so höher, je geringer die Anzahl der zur Auswahl stehenden Publikationen ist. Je kleiner die Gruppe ist, desto höher ist die Sichtbarkeit der einzelnen Publikation.

6.2.2.4 Die Nutzung von Einzelpublikationen

Aus den vorangegangenen Abschnitten ergeben sich mögliche Erklärungen, warum Einzelpublikationen im Vergleich zu den anderen Publikationstypen geringere Downloads aufweisen, wenn Rezipienten die Browsing-Homepage als Einstiegsseite benutzen und von dort aus den Inhalt des IR erkunden.

Obwohl sich die Navigation durch die OPUS-IR ehsStu, HeiDok und SciDok von der von edoc stark unterscheidet, sind bezüglich der Einzelpublikationen doch Gemeinsamkeiten vorhanden. Bei den OPUS-IR befindet sich der Einstieg zu einem großen Teil der Einzelpublikationen (Schriftenreihen, Portale und Collections) im unteren Teil einer Liste. Von dort aus gibt es noch einen weiteren Link, um an die Publikationen selbst zu gelangen. Man kann zwar auch über einen Link „Dokumentarten“ zu den Einzelpublikationen navigieren, aber unter den Dokumentarten hat von den Einzelpublikationen nur „Aufsatz“ eine günstige Position.

Die Anzahl der Publikationen ist aber wiederum relativ groß, so dass der Vorteil der Position verlorengeht. Bei edoc liegt die Einstiegsmöglichkeit zu Einzelpublikationen generell hinter Qualifikationsarbeiten und man gelangt bis auf Ausnahmen über noch mehr Ebenen als bei den OPUS-IR zur Publikation. Die Einzelpublikationen sind also durch die Position im Browsing noch mehr als bei OPUS benachteiligt. Des Weiteren ist es nicht möglich, durch Navigation zu den Einzelpublikationen von bestimmten Autoren anders als über eine weitere Ebene zu gelangen.

Mit Hilfe von NoRA könnte geprüft werden, ob sich durch die Verbesserung der Position im Browsing und durch Verringerung der Ebenen, über die der Rezipient zu Einzelpublikationen gelangt, deren Nutzung verbessern lässt.

6.2.2.5 Schlussfolgerungen

Wie genau Downloads von Publikationen von der Position im Browsing, der Größe der Gruppen und weiteren äußeren Faktoren abhängen, muss mit geeigneten Testanordnungen untersucht werden. Dass solche Einflüsse existieren, ist aber offensichtlich. Wenn die Unterschiede in den Downloads nicht maßgeblich durch den Inhalt, sondern z. B. durch die Position der Publikationen im Browsing erklärt werden können, liegt die Vermutung nahe, dass Nutzer von IR häufig über die Homepage zu den Publikationen gelangen. Sie versuchen, einen Überblick zu bekommen, wie sich die Universität im Spiegel der Publikationen auf dem IR darstellt. Der Auslöser für einen Download ist häufig nicht das Interesse für einen bestimmten Inhalt, sondern das Interesse daran, wer veröffentlicht und was diejenige Person veröffentlicht. Dabei ist die Wahrscheinlichkeit groß, dass ein zufällig an oberer Position auf der Website zu findender Link gewählt wird. Ist die folgende Auswahl an Publikationen hinter diesem Link klein, steigt die Chance auf einen Download für jede einzelne davon. Im Vordergrund steht auch oft das Interesse an Publikationen von bestimmten Personen.

Die Aufgabe der Betreiber sollte sein, die oben dargestellten Verzerrungseffekte so gering wie möglich zu halten. Daraus ergeben sich eine Reihe von Konsequenzen für die Gestaltung des Browsings. Der völlige Verzicht und die Reduzierung auf eine Suche als Einstiegsmöglichkeit würde zwar Verzerrungen weitgehend vermeiden, kann aber nicht die Lösung sein, da Nutzer von IR eine Homepage benötigen, von der aus sich der Inhalt des Repository erschließen lässt. Jedoch sollten Sichtbarkeitsvorteile aufgrund der Auffindbarkeit von der Homepage aus so gering wie möglich sein. Mit Hilfe von NoRA kann festgestellt werden, ob sich Veränderungen im Browsing in beabsichtigter Weise auf die Nutzung auswirken.

Vor allem Herausgeber von Zeitschriften und anderer Periodika sind natürlich daran interessiert, die Inhalte in einer von ihnen gewünschten Form zu veröffentlichen. Hier müssen Kompromisse gefunden werden. Andererseits kann durch die Gestaltung der Website die Sichtbarkeit gezielt verbessert werden, was besonders für die Einzelpublikationen in Betracht gezogen werden sollte. Bei all dem muss die Sichtbarkeit des gesamten IR und nicht die Sichtbarkeit einzelner Gruppen im Vordergrund stehen. So wäre es nicht zielführend, die Akquisition von Publikationen kleiner Gruppen nicht zu verstärken, nur weil sich dadurch die Downloads einzelner Publikationen verringern könnten.

Für die Gestaltung des IR ist es wichtig, die Veränderungen im Bestand an Publikationen zu beobachten und darauf zu reagieren. Wie sich gezeigt hat, orientiert sich bei allen vier IR das Browsing immer noch zu sehr an den Hochschulschriften, obwohl die Entwicklung längst eine andere Richtung nimmt. Einzelpublikationen als Form des Institutional Self Archiving und besonders Periodika müssen im Gesamtzusammenhang der IR aufgewertet werden. In diesem Sinne sind auch die Hinweise zu verstehen, die im Kapitel „Ergebnisse“ für die einzelnen IR gegeben werden.

Die Ergebnisse der Analysen zeigen, welche Faktoren die Downloads von Publikationen beeinflussen. Das hat nicht nur Konsequenzen für die Gestaltung des IR, sondern auch für die Bewertung der Downloads, die für die einzelne Publikation erhoben werden. Finden Autor oder Rezipient zusammen mit der Publikation eine Downloadstatistik vor, gehen sie verständlicherweise davon aus, dass die Zahlen widerspiegeln, wie oft die Publikation genutzt wurde und sehen die Downloads als ein Maß für das Interesse an. Dass die angegebenen Downloadzahlen nicht nur durch menschliche, sondern auch maschinelle Zugriffe entstehen und unter unterschiedlichen Bedingungen ermittelt werden, ist erkannt worden. Downloads, so wie sie zurzeit von den IR erhoben werden, gelten nicht als geeignet, um Vergleiche zwischen Publikationen verschiedener IR anzustellen. Diesem Problem soll durch standardisierte, zentral erstellte Nutzungsstatistiken begegnet werden. In Zukunft wird es möglich sein, von einem gemeinsamen Portal aus auf die Publikation vieler OA Repositories zugreifen zu können. Dies ist das Ziel einer Reihe von nationalen und internationalen Projekten (siehe Abschnitt 1.4 Institutional Repositories). So soll ein zentraler Nachweis aller OA-Publikationen Deutschlands geschaffen werden. Dabei werden dem Nutzer zusammen mit der Publikation auch die Ergebnisse der standardisierten Downloadstatistik als Mehrwert zur Verfügung gestellt. Dubletten sollen ermittelt und Downloads zusammengefasst werden. Damit käme es zum Vergleich von Downloadstatistiken, die zwar unter gleichen Bedingungen errechnet wurden, aber durch die Sichtbarkeit auf den IR beeinflusst sind. Dieser Aspekt findet in der gegenwärtigen Diskussion um Nutzungszahlen bisher keine Aufmerksamkeit. Ein Grund dafür ist die Annahme, dass nur sehr wenige Nutzer die Publikationen über die Homepage des IR finden, sondern dass die Nutzung von Suchmaschinen dominiert. Dagegen spricht allerdings, dass Unterschiede in den Downloads bei den vier untersuchten IR gut durch Sichtbarkeitseffekte, die durch das Browsing entstehen, erklärt werden können. Ein gemeinsames Portal vieler oder aller deutscher OA-Repositories würde solchen Effekten entgegenwirken, sie aber nicht völlig beseitigen.

Vor dem Hintergrund der Kritik an der Verwendung von Maßen wie dem Citation Impact als Kriterium zur Bewertung von wissenschaftlichen Publikationen und deren Autoren wird deutlich, dass im Umgang mit Downloads als Impact-Maß noch viel sensibler umgegangen werden muss. Aufgrund der Ergebnisse der Nutzungsdatenanalyse von vier OA Repositories muss erneut hinterfragt werden, was Downloads aussagen können. Für Publikationsgruppen geben sie Auskunft über die Sichtbarkeit, möglicherweise auch über das Interesse an ihrem Inhalt, kaum aber über Qualität. Darüber hinaus können Robot-Zugriffe nie vollständig aus der Statistik eliminiert werden, so dass die Angabe von Downloads für die einzelne Publikation immer unter diesem Vorbehalt zu sehen ist. Um bei Nutzern, Herausgebern und Autoren keinen falschen Eindruck entstehen zu lassen, muss bei der Veröffentlichung von Downloadstatistiken unbedingt auf diesen Umstand hingewiesen werden.

Zusammenfassung und Ausblick

Die Methode ordnet sich in den Gesamtzusammenhang des wissenschaftlichen Publizierens ein. Im Mittelpunkt der vorgelegten Arbeit steht die Entwicklung der Analysemethode NoRA, mit der Metadaten und Nutzungsdaten von Institutional Repositories (IR) analysiert werden können.

Im 1. Kapitel werden die Fragen und Probleme behandelt, welche den Ausgangspunkt der Arbeit bilden und zu ihrer Aufgabenstellung führen. Das zentrale Thema ist der Impact wissenschaftlicher Publikationen im Allgemeinen und im Besonderen der Impact elektronischer Publikationen, die nach dem Prinzip des Open Access (OA) auf IR veröffentlicht werden. Nutzungsdaten elektronischer Publikationen finden vor allem im Zusammenhang mit der OA-Bewegung große Beachtung. Autoren und Herausgeber, die auf IR publizieren, gehen davon aus, dass Downloads ein Anzeichen für das Interesse an der Publikation und deren Verbreitung sind. Den Download Impact elektronischer Publikationen von IR zu steigern, ist das Ziel, welches Betreiber von IR im Interesse ihrer Autoren und Herausgeber verfolgen. Die Sichtbarkeit der OA-Publikationen im Internet ist die Voraussetzung für die Erreichung dieses Zieles. Welche Schlussfolgerungen aus der Höhe des Download Impact wirklich gezogen werden können, ist Gegenstand der aktuellen Forschung. Eine entscheidende Frage ist, wie die Sichtbarkeit eines IR und damit seiner Publikationen nachhaltig verbessert werden kann. Um hierzu verlässliche Aussagen zu erhalten, ist es notwendig, die gegenwärtige Nutzung von IR zu analysieren. Solch eine Analyse bietet die Methode NoRA. Mit Hilfe von NoRA werden Gruppen von Publikationen erkannt, die sich im Hinblick auf ihre Nutzung unterscheiden, und der Unterschied quantitativ erfasst.

Wie die Zugriffsdaten aus den Logfiles von Webservern ermittelt werden und von welchen Faktoren die Anzahl der Downloads abhängt, wird im 2. Kapitel ausführlich behandelt. Anhand von Beispielen wird gezeigt, dass die Höhe der Downloads maßgeblich davon abhängt, welche Programme und Algorithmen für deren Ermittlung verwendet und welche Robot-Zugriffe ausgeschlossen werden. Das führt dazu, dass für einzelne Publikationen die Nutzungsstatistik unzuverlässig ist. Ein Vergleich von absoluten Nutzungsdaten verschiedener IR ist generell nicht sinnvoll. Dem soll mit standardisierter Erfassung und zentraler Aggregation der Nutzungsdaten von IR, die in einer gemeinsamen Infrastruktur zusammengeschlossen sind, begegnet werden. Die Projekte, welche sich mit dieser Problematik befassen, befinden sich zurzeit noch im Entwicklungsstadium. Trotz der bekannten Mängel der ermittelten Downloads ist es jedoch möglich, wertvolle Informationen aus diesen Daten zu gewinnen. Berücksichtigt man ihre statistischen Eigenschaften, können durch Anwendung adäquater Verfahren vergleichende Aussagen über die Nutzung von Publikationsgruppen eines IR getroffen werden.

Die statistischen Eigenschaften der Downloads werden eingehend untersucht und in Form von Grafiken dargestellt. Es wird begründet, warum nichtparametrische Verfahren der Mathematischen Statistik geeignet sind, Downloads von IR zu analysieren. Anschließend wird dargestellt, wie auf der Grundlage dieser Verfahren die Analysemethode NoRA entwickelt wurde, mit der es möglich ist, aus den Publikationen eines IR Gruppen zu identifizieren, deren Downloads sich signifikant unterscheiden. Der Vergleich dieser Publikationsgruppen ist

anhand ihrer Mediane möglich. Die Mediane informieren über graduelle Unterschiede in der Sichtbarkeit von Publikationsgruppen. Zusammen mit der Analyse der Metadaten erhalten Betreiber Hinweise für die strukturelle Verbesserung des IR. Der Prozess der Analyse wird so weit formalisiert, dass alle notwendigen Schritte von der Vorbereitung und Auswahl des Datenmaterials über die Durchführung der statistischen Verfahren bis zur grafischen Darstellung der Ergebnisse für Betreiber von IR nachvollziehbar und ohne Spezialkenntnisse durchführbar sind.

Anschließend wird die praktische Anwendung am Beispiel der Daten von vier IR demonstriert. Alle der insgesamt 6 Schritte der Analyse werden nacheinander durchlaufen und dabei gezeigt, wie einerseits mit vorgegebenen Kategorien von formalen und inhaltlichen Merkmalen der Publikationen Gruppen gebildet werden können, andererseits aber auch die Berücksichtigung spezifischer Merkmale der IR möglich ist. Die Beispiele können von den Betreibern von IR mit eigenem Datenmaterial nachvollzogen zu werden. Sie zeigen neben der Berücksichtigung vorgegebener Kategorien von Merkmalen auch die flexiblen Anwendungsmöglichkeiten von NoRA für die Analyse von Downloads. Die Ergebnisse für alle vier IR werden dokumentiert und interpretiert. In allen Fällen weisen die Analyseergebnisse von NoRA auf Verbesserungsmöglichkeiten bei der Sichtbarkeit von Publikationen und damit der Qualität des IR hin.

Beim Vergleich der Ergebnisse der vier IR wird deutlich, dass es Zusammenhänge zwischen der Struktur des Webauftritts der IR und der Nutzung von Publikationsgruppen gibt, die in dieser Form nicht bekannt waren. Im letzten Kapitel wird außerdem dargelegt, wie die Analyseergebnisse von vier IR interessante Einsichten in das Verhalten der Nutzer von IR geben. Beide Aspekte werden in der aktuellen Diskussion um Nutzungsdaten zurzeit nicht berücksichtigt. Für die Betreiber von IR ergeben sich daraus jedoch eine Reihe von Konsequenzen, welche bisher nicht abzusehen waren. Die systematische Untersuchung der Einflüsse durch die Gestaltung der Websites von IR auf die Nutzung von Publikationen, die in Zukunft stärkere Beachtung finden müssen, bietet ein breites und interessantes Anwendungsfeld für NoRA.

Im Hinblick auf den künftigen Umgang mit Nutzungsdaten sind weitere Überlegungen angebracht, da die mithilfe von NoRA gewonnen Erkenntnisse über die Nutzung elektronischer Publikationen von IR Fragen über den Zweck der Erfassung und die Interpretation von Nutzungsstatistiken aufwerfen. Vor dem Hintergrund der vielfältigen Einflüsse auf die Downloads muss die Veröffentlichung einer Nutzungsstatistik auf Publikationsebene kritisch bewertet werden. Ein Hinweis für die Rezipienten der Publikationen, dass die Angaben der Nutzungsstatistik nicht zuverlässig sind und keinen Vergleich mit anderen Publikationen zulassen, ist erforderlich. Außerdem sollten Autoren und Herausgeber selbst entscheiden können, ob eine Nutzungsstatistik veröffentlicht wird, da nicht auszuschließen ist, dass aus der Höhe der Downloads falsche Schlussfolgerungen gezogen werden. Wie im 1. Kapitel dargelegt wird, sind Zitationen nicht generell als Belohnung für wissenschaftliche Qualität zu sehen und darauf beruhende Impact-Maße erfordern Sorgfalt bei der Interpretation. Das ist einer der Gründe dafür, dass der Citation Impact nur zusammen mit anderen Kriterien zur Bewertung wissenschaftlicher Leistung und Qualität herangezogen werden kann. Noch weniger trifft die Annahme zu, dass Downloads von OA-Publikationen generell ein Interesse am Inhalt ausdrücken. Mit

Nutzungsstatistiken und damit dem Download Impact von einzelnen Publikationen muss demzufolge äußerst sorgsam umgegangen werden.

Wie die Anwendung von NoRA zeigt, kann die Analyse von Downloads jedoch wertvolle Informationen über die Nutzungshäufigkeit von Publikationsgruppen und das Verhalten der Nutzer liefern, wenn deren statistische Eigenschaften durch die Verwendung geeigneter Verfahren berücksichtigt werden. Wird das Konzept von NoRA auf ein Netzwerk von IR übertragen, deren Nutzungsdaten unter standardisierten Bedingungen erfasst und zentral aggregiert werden, kann theoretisch nach dem gleichen Prinzip die Nutzung von IR analysiert werden. Unter der Voraussetzung eines solchen Netzwerkes, dessen Entwicklung das Ziel mehrerer Forschungsprojekte ist, wird es möglich sein, Downloads der beteiligten IR zu vergleichen. Es wären ähnliche Untersuchungen der Downloads von IR notwendig, die zur Entwicklung von NoRA geführt haben, um konkrete Rahmenbedingungen für die Datenauswahl festzulegen. Prinzipiell steht einer Anwendung der gleichen statistischen Verfahren jedoch nichts im Weg und ein „Repository Impact Factor“, der die Nutzung und damit die Sichtbarkeit von IR einbezieht, könnte definiert werden.

Literaturverzeichnis

Adler, Robert; Ewing, John und Taylor, Peter (2008): Citation Statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS), http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf [28.12.08].

Bartneck, Christoph und Kokkelmans, Servaas (2010): Detecting h-index manipulation through self-citation analysis, *Scientometrics* 87 [1], S. 85-98.

Björk, Bo-Christer; Welling, Patrik; Laakso, Mikael; Majlender, Peter; Hedlund, Turi und Guðnason, Guðni (2010): Open Access to the Scientific Journal Literature: Situation 2009, PLoS ONE 5, http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0011273#top [20.5.2011].

Bollen, Johan; Luce, Rick; Vemulapalli, Soma Sekhara und Xu, Weining (2003): Usage Analysis for the Identification of Research Trends in Digital Libraries, D-Lib Magazine 9, http://www.dlib.org/dlib/may03/bollen/05bollen.html [20.5.2011].

Bollen, Johan; Rodriquez, Marko A. und Sompel, Herbert Van de (2006): Journal Status, *Scientometrics* 69 [3], S. 669-687.

Bollen, Johann und van de Sompel, Herbert (2008): Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics, *Journal of the American Society for Information Science and Technology* 59 [1], S. 136–149.

Bomhardt, Christian; Gaul, Wolfgang und Schmidt-Thieme, Lars (2005): Web Robot Detection - Preprocessing Web Logfiles for Robot Detection, New Developments in Classification and Data Analysis, Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, September 22–24, 2003, S. 113-124.

Bornmann, Lutz; Wallon, Gerlind und Ledin, Anna (2008): Is the h index related to (standard) bibliometric measures and to the assessments by peers? An investigation of the h index by using molecular life sciences data *Research Evaluation* 17 [2], S. 149-156.

Braun, Tibor; Glänzel, Wolfgang und Schubert, András (2006): A Hirsch-type index for journals, *Scientometrics* 69 [1], S. 169-173.

Brin, Sergej und Page, Larry (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of the 7th international conference on World Wide Web (WWW). April 14-18, Brisbane, Australia., S. 107–117.

Brody, Tim; Harnad, Stevan und Carr, Leslie (2006): Earlier Web Usage Statistics as Predictors of Later Citation Impact, *Journal of the American Association for Information Science and Technology (JASIST)* 57(8), S. 1060-1072.

COUNTER (2006): Release 1 of the COUNTER Code of Practice for Books and Reference Works, http://www.projectcounter.org/cop/books/cop_books_ref.pdf [20.5.2011].

COUNTER (2008): Release 3 of the COUNTER Code of Practice for Journals and Databases, http://www.projectcounter.org/r3/Release3D9.pdf [20.5.2011].

Cozzens, Susan E. (1989): What do citations count? the rhetoric-first model *Scientometrics* 15 [5-6], S. 437-447.

Craig, Iain D.; Plume, Andrew M.; McVeigh, Marie E.; Pringle, James und Amin, Mayur (2007): Do Open Access Articles Have Greater Citation Impact? A critical Review ofthe literature, *Journal of Infometrics* 1 [3], S. 239-248.

Cronin, Blaise (2001): Bibliometrics and beyond: some thoughts on web-based citation analysis, *Journal of Information Sciece* 27, 1, S. 1-7.

Crow, Raym (2002): The Case for Institutional Repositories: A SPARC Position Paper, http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf [3.1.07].

Darmoni, Stefan J.; Roussel, Francis; Benichou, Jacques; Thirion, Benoit und Pinhas, Nicole (2002): Reading factor: a new bibliometric criterion for managing digital libraries, *Journal of the Medical Library Association* 90 [3], S. 323–327.

Davis, Philip M.; Lewenstein, Bruce V.; Simon, Daniel H.; Booth, James G. und Conolly, Matthew J. L. (2008): Open access publishing, article downloads, and citations: randomised controlled trial, *British Medical Journal* 337:a568.

de Solla Price, Derek (1965): Networks of Scientific Papers, *Science* 149 [3683], S. 510-515.

de Solla Price, Derek (1974): Little Science, Big Science. Von der Studierstube zur Großforschung, Suhrkamp Taschenbuch Wissenschaft 48, ISBN: 3-518-07648-5.

Dettmar, Gebhard (2004): Community of Knowledge, Knowledge Discovery in Databases, Teil III: Konzept Hierarchien in WUMprep, http://www.c-o-k.de/cp_artikel.htm?artikel_id=158 [16.2.2008].

DINI (2005): Elektronisches Publizieren an Hochschulen: Inhaltliche Gestaltung der OAI-Schnittstelle - Empfehlungen -, DINI-Schriften 2, (urn:nbn:de:kobv:11-10049220) [28.6.2010].

DINI (2010a): DINI-Zertifikat Dokumenten- und Publikationsservice 2010, DINI-Schriften 3, (urn:nbn:de:kobv:11-100109986) [4.1.2011].

DINI (2010b): Gemeinsames Vokabular für Publikations- und Dokumenttypen, DINI-Schriften 12, (urn:nbn:de:kobv:11-100109998) [4.1.2011].

Dobratz, Susanne und Müller, Uwe Thomas (2009): Wie entsteht ein Institutional Repository? – Eine systematische Hinführung in acht Schritten, cms-journal 32, (urn:nbn:de:kobv:11-10098215) [5.1.2011].

Dobrov, Gennadij M. (1974): Wissenschaft: ihre Analyse und Prognose, Stuttgart: DVA, ISBN: 3-421-02266-6.

Drèze, Xavier und Zufryden, Fred (2004): Measurement of online visibility and its impact on Internet traffic, *Journal of Interactive Marketing* 18 [1], S. 20–37.

Egghe, Leo (2006): Theory and practise of the g-index, *Scientometrics* 69 [1], S. 131-152.

Eichmann, David (1995): Ethical Web agents, *Computer Networks and ISDN Systems* 28 (1-2), S. 127-136.

Foster, Nancy F. und Gibbons, Susan (2005): Understanding faculty to improve content recruitment for institutional repositories, D-Lib Magazine 11, http://www.dlib.org/dlib/january05/foster/01foster.html [22.02.2007].

Garfield, Eugene (1955): Citation indexes to science: a new dimension in documentation through association of ideas, *Science* 122(3159), S. 108-111.

Garfield, Eugene (2005): The Agony and the Ecstasy-The History and Meaning of the Journal Impact Factor, International Congress on Peer Review and Biomedical Publication, http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf [19.1.07].

Geens, Nick; Huysmans, Johan und Vanthienen, Jan (2006): Evaluation of Web Robot Discovery Techniques: A Benchmarking Study, Advances in Data Mining, 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006. Proceedings, Buchreihen Lecture Notes in Computer Science, Volume 4065/2006, S. 121-130.

Gonzalez-Pereira, Borja; Guerrero-Bote, Vicente und Moya-Anegon, Felix (2010): The SJR indicator: A new indicator of journals' scientific prestige, preprint: http://arxiv.org/abs/0912.4141 [10.1.2011].

Haque, Asif-ul und Ginsparg, Paul (2009): Positional effects on citation and readership in arXiv, *Journal of the American Society for Information Science and Technology* 60 [11], S. 2203–2218.

Harnad, Stevan (2006): Preprints, Postprints, Peer Review, and Institutional vs. Central Self-Archiving, Open Access Archivangelism, http://openaccess.eprints.org/index.php?/archives/140-Preprints,-Postprints,-Peer-Review,-and-Institutional-vs.-Central-Self-Archiving.html [14.1.2011].

Harnad, Stevan; Brody, Tim; Vallières, Francois; Carr, Les; Hitchcock, Steve; Gingras, Yves; Oppenheim, Charles; Stamerjohanns, Heinrich und Hilf, Eberhard R. (2004): The Access/Impact Problem and the Green and Gold Roads to Open Access, *Serials Review* 30 [4], S. 310-314.

Havemann, Frank (2009): Einführung in die Bibliometrie, Gesellschaft für Wissenschaftsforschung, Berlin, ISBN: 978-3-934682-46-7.

Heindl, Eduard (2003): Logfiles richtig nutzen, Galileo Press, Bonn, ISBN: 9783898424011.

Hess, Thomas; Wigand, Rolf T.; Mann, Florian und von Walter, Benedikt (2007): Open Access & Science Publishing - Results of a Study on Researchers' Acceptance and Use of Open Access Publishing, Management Reports of the Institute for Information Systems and New Media, LMU München, Munich, Nr. 1/07, http://openaccess-study.com/Hess_Wigand_Mann_Walter_2007_Open_Access_Management_Report.pdf [24.1.2011].

Hirsch, Jorge E. (2005): An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences,* 102 [46], S. 16569-16572.

Hornbostel, Stefan (1997): Wissenschaftsindikatoren: Bewertungen in der Wissenschaft, Westdeutscher Verlag, Opladen, ISBN: 3-531-12908-2.

Ingwersen, Peter (1998): The calculation of Web Impact Factors, *Journal of Documentation* 54 [2], S. 236-243.

Jin, Bihui; Liang, Liming; Rousseau, Ronald und Egghe, Leo (2007): The R- and AR-indices: Complementing the h-index, *Chinese Science Bulletin* 52 [6], S. 855-863.

Kähler, Wolf-Michael (2010): Statistische Datenanalyse. Verfahren verstehen und mit SPSS gekonnt einsetzen, Vieweg+Teubner Verlag | Springer Fachmedien Wiesbaden GmbH, ISBN: 978-3-8348-1326-8.

Kaplan, Nancy R. und Nelson, Michael L. (2000): Determining the publication impact of a digital library, *Journal of the American Society for Information Science and Technology*, S. 324–339.

Katz, Leo (1953): A new status index derived from sociometric analysis, *Psychometrika* 18 [1], S. 39-43.

Kaufmann, Noogie C. (2007): Speicherung von IP-Adressen auf Websites verboten ?, DFN Infobrief Recht, http://www.dfn.de/fileadmin/3Beratung/Recht/1infobriefearchiv/07-09/Infobrief_10_07.pdf [24.1.2011].

Kurtz, Michael J.; Eichhorn, Guenther; Accomazzi, Alberto; Grant, Carolyn; Demleitner, Markus und Murray, Stephen S. (2005a): Worldwide use and impact of the NASA Astrophysics Data System digital library (published online 2004), *Journal of the American Society for Information Science and Technology* 56 [1], S. 36-45.

Kurtz, Michael J.; Eichhorn, Guenther; Accomazzi, Alberto; Grant, Carolyn; Demleitner, Markus; Murray, Stephen S.; Martimbeau, Nathalie und Elwell, Barbara (2005b): The bibliometric properties of article readership information (published online 2004), *Journal of the American Society for Information Science and Technology* 56 [2], S. 111-128.

Kurtz, Michael J. und Henneken, Edwin A. (2007): Open Access does not increase citations for research articles from The Astrophysical Journal, http://arxiv.org/abs/0709.0896 [24.1.2011].

Lawrence, S (2001): Free online availability substantially increases a papers impact, Nature 411, http://www.nature.com/nature/journal/v411/n6837/full/411521a0.html [18.5.2011].

Mabe, Michael (2003): The growth and number of journals, *The Journal for the Serials Community* 16 [2], S. 191 - 197.

Malitz, Robin (2009): Open Access – Verfügbar ist noch nicht präsent – Verbesserung der Sichtbarkeit von Open Access Repositories durch die Bildung von Netzwerken, cms-journal 32, (urn:nbn:de:kobv:11-10098276) [15.3.2011].

Mann, Florian; von Walter, Benedikt; Hess, Thomas und Wigand, Rolf T. (2009): Open Access Publishing in Science, Communications of the ACM 52, http://cacm.acm.org/magazines/2009/3/21798-open-access-publishing-in-science/fulltext [24.1.2011].

Mayr, Philipp (2009): Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in digitalen Bibliotheken, Dissertation, Humboldt-Universität zu Berlin, Philosophische Fakultät I, (urn:nbn:de:kobv:11-10097590) [25.1.2011].

Merton, Robert K. (1968): The Matthew Effect in Science, *Science* 159 [3810], S. 56-63.

Merton, Robert K. (1972): Wissenschaft und demokratische Sozialstruktur, In Weingart, P. (Hrsg.): Wissenschaftssoziologie, Frankfurt/M., S. 45-59.

Michailow, Aleksandr I.; Cernyi, Arkadij I. und Giljarevskij, Rudzero S. (1970): Grundlagen der wissenschaftlichen Dokumentation und Information, Westdeutscher Verlag, Köln.

Moed, Henk F. (2010): Measuring contextual citation impact of scientific journals, *Journal of Informetrics* 4 [3], S. 265-277.

Müller, Uwe Thomas (2009): Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – systematische Klassifikation und empirische Untersuchung, Dissertation, Humboldt-Universität zu Berlin, Philosophische Fakultät I, (urn:nbn:de:kobv:11-10096430) [1.3.2011].

Noruzi, Alireza (2006): The Web Impact Factor: A Critical Review, The Electronic Library 24, http://eprints.rclis.org/5543/ [2.12.2008].

Odlyzko, Andrew (2002): The rapid evolution of scholarly communication, *Learned Publishing* 15 [1], S. 7-19.

Ohly, H. Peter (2010): Zitationsanalyse: Beschreibung und Evaluation von Wissenschaft, In Stegbauer, Christian und Häußling, Roger (Hrsg.): Handbuch Netzwerkforschung, VS Verlag für Sozialwissenschaften, S. 785-797.

Pritchard, Alan (1969): Statistical bibliography or bibliometrics, *Journal of Documentation* 25 [4], S. 348-349.

Schirmbacher, Peter (2005): Die neue Kultur des elektronischen Publizierens, cms-journal 27, (urn:nbn:de:kobv:11-10044162) [21.1.2010].

Schmidt-Mänz, Nadine und Gaul, Wolfgang (2004): Web Mining and Online Visibility, Classification - the Ubiquitous Challenge, Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, March 9-11, 2004, Series: Studies in Classification, Data Analysis, and Knowledge Organization, S. 418-425.

Severiens, Thomas und Hilf, Eberhard R. (2004): Elf Argumente für Open Access http://www.isn-oldenburg.de/publications/11argumente.html [4.1.2010].

SPARC (2002): SPARC Institutional Repository Checklist & Resource Guide, http://www.arl.org/sparc/bm~doc/ir_guide__checklist_v1.pdf [1.5.2011].

Spiliopoulou, Myra und Faulstich, Lukas C. (1999): WUM: A Tool for Web Utilization Analysis, *Lecture Notes in Computer Science* 1590, S. 184-203.

Sponsler, Ed (2004): An Eprints Apache Log Filter for Non-Redundant Document Downloads by Browser Agents, http://caltechlib.library.caltech.edu/73/01/Report-2004-NOV.pdf [14.2.2008].

Stassopouloua, Athena und Dikaiakos, Marios D. (2009): Web robot detection: A probabilistic reasoning approach, *Computer Networks* 53 [3], S. 265-278.

Stock, Wolfgang G. (1998): Was ist eine Publikation? Zum Problem der Einheitenbildung in der Wissenschaftsforschung, In Fuchs-Kittowski, Klaus; Laitko, Hubert; Parthey, Heinrich und Umstätter, Walter (Hrsg.): Wissenschaftsforschung Jahrbuch 1998, S. 239-282.

Suber, Peter (2004): Praising progress, preserving precision, SPARC Open Access Newsletter, http://www.earlham.edu/~peters/fos/newsletter/09-02-04.htm [13.5.08].

Swan, Alma (2006a): The culture of open access: researchers' views and responses, In Jacobs, Neil (Hrsg.): Open Access: Key Strategic, Technical and Economic Aspects, Chandos Publishing, Oxford, S. 65-72.

Swan, Alma (2006b): Overview of scholarly communication, In Jacobs, Neil (Hrsg.): Open Access: Key Strategic, Technical and Economic Aspects, Chandos Publishing, Oxford, S. 3-12.

Swan, Alma (2010): The Open Access citation advantage: Studies and results to date, http://eprints.ecs.soton.ac.uk/18516/ [10.3.2011].

UNESCO (2007): Open Access. Chancen und Herausforderungen– ein Handbuch, Deutsche UNESCO-Kommission, ISBN: 3-927907-96-0

Vaas, Rüdiger (2010): Freie Wissenschaft für alle. Bild der Wissenschaft online 3, http://www.bild-der-wissenschaft.de/bdw/bdwlive/heftarchiv/index2.php?object_id=32192855 [9.1.2011].

Walker, Dylan; Xie, Huafeng; Yan, Koon-Kiu und Maslov, Sergei (2007): Ranking scientific publications using a model of network traffic, Journal of Statistical Mechanics: Theory and Experiment, (http://iopscience.iop.org/1742-5468/2007/06/P06010) [1.5.2011].

Weingart, Peter (2003): Wissenschaftssoziologie, Reihe Einsichten, transcript Verlag Bielefeld, ISBN: 3-933127-37-8.

Weishaupt, Karin (2009): Open-Access-Zeitschriften – Entwicklung von Maßnahmen zur Akzeptanzsteigerung auf der Basis einer Autorenbefragung, Dissertation Humboldt-Universität zu Berlin, Philosophische Fakultät I, (urn:nbn:de:kobv:11-100100107) [1.5.2011].

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ARPANET	Advanced Research Projects Agency Network
ASCII	America Standard of Information Interchange
CLF	Common Logfile Format
COUNTER	Counting Online Usage Of NeTworked Electronic Resources
CSS	Cascading Style Sheet
DDC	Dewey Decimal Classification
DINI	Deutsche Initiative für NetzwerkInformation
DNS	Domain Name Server
DRIVER	Digital Repository Infrastructure Vision for European Research
ECLF	Extended Common Logfile Format
FTP	File Transfer Protocol
GMT	Greenwich Mean Time
GPL	General Public License
HTTP	Hyper Text Transfer Protocol
ICOLC	International Coalition of Library Consortia
IFABC	International Federation of Audit Bureaux of Circulations
IIS	Microsoft Internet Information Server
IMU	International Mathematical Union
IOCLC	International Coalition of Library Consortia
IP	Internet Protocol
IR	Institutional Repository
ISI	Institute for Scientific Information, frühere Bezeichnung für Thomson Reuters
JIF	Journal Impact Factor
LANL	Los Alamos National Library
NCSA	National Center for Supercomputing Applications
NoRA	Non-parametric Repository Aanalysis
OA	Open Access
OAI	Open Archive Initiative
ODBC	Open Database Connectivity
OPUS	Online Publikationsverbund Universität Stuttgart
PASW	Predictive Analysis SoftWare
PDF	Portable Document Format
PIRUS	Publisher and Institutional Repository Usage Statistics
PMH	Protocol for Metadata Harvesting
PS	Post Script
RoMEO	Rights MEtadata for Open archiving

SAS	Statistical Analysis Software
SCI	Science Citation Index
SHERPA	Securing a Hybrid Environment for Research Preservation and Access
SNIP	Source Normalized Impact per Paper
SPARC	Scholarly Publishing Academic Resource Coalition
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SRJ	SCImago Journal Rank
SURE	Statistics on the Usage of Repositories
SUSHI	Standardised Usage Statistics Harvesting Initiative
TA	Toll Access
TXT	Erweiterung für Text-Datei
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WIF	Web Impact Factor
WoS	Web of Science
WUM	Web Utilization Miner
XHTML	Extended HyperText Markup Language
XML	Extended Markup Language

Anhang A: Anschreiben und Datenbeschreibung

Dear Repository Manager,

My name is Sabine Henneberger. I am a member of the Electronic Publishing Group of the Computer and Media Services at the Humboldt University Berlin. In addition to my regular work in this department, I focus on potential improvements of repository structure in my doctoral thesis. Part of my thesis aims at understanding the usage of open access repositories. My work is supervised by Peter Schirmbacher who is a professor for Information Management and director of the Computer and Media Service at the Humboldt-University.

Download statistics represent the data sets on which I work. Unlike studies that focus on individual documents (such as the project Open-Access-Statistik http://www.dini.de/projekte/oa-statistik/english/) I am interested in the download statistics of the entire repository. The aim is to compare download numbers of document groups and to draw conclusions for further development of the content of the repository.

So far, four German repositories participate in the study. The documents were classified into groups of subjects and publication types. Interestingly, it turned out that groups with the largest increase in numbers over time have the lowest download numbers. To increase impact and visibility of the entire repository it would be helpful to acquire more content of certain other groups of documents.

Now I would also like to analyze download data of additional open access repositories to verify the method and to check whether the findings can be further generalized. It would be most interesting to analyze not only German but also foreign repositories. Therefore, I would like to ask you if it were possible to get data from a repository you are working with. It certainly takes time and effort to produce such data. However, I am sure this is worthwhile since conclusions drawn from this analysis will provide helpful hints for future extension of content.

I would be happy if you decided to participate in this study. If you do so, may I kindly ask you to provide me with information about the repository of your and how the document specific download data was generated (please see table 1 of the attachment). The document specific download data itself would need to be structured in a specific way to ensure coherence of the study (table 2 Meta Data, table 3 Download Data).

The classification of meta data Subject Type and Content Type is based on the terminology of OpenDOAR, but also any other classification is welcome. If providing the complete meta data is not possible, I would also be interested in subsets. Subject Type and Content Type are more important than Language, Author or Publisher. Monthly download data are preferred, but I can also analyze annual or otherwise aggregated data. I can also use shorter or non-contiguous periods.

The data can be stored in one single table, covering meta data and download data, or in several tables for meta data and download data as well. Text files are generally sufficient, but formats like SAS, SPSS or EXCEL are also welcome.

I hope I could interest you in my study. Please contact me in case of questions or suggestions by e-mail or telephone. Also, if you know of any other potential participants who might be interested in this kind of study and would be able to provide similar data, I were grateful if you either forwarded this email or provided me with contact details.

Yours sincerely,

Sabine Henneberger

Sabine Henneberger, Humboldt-Universität zu Berlin, 02/02/2010

Data for Repository Statistics

1 Informations about the Repository and Statistics

1.1 Information about the Repository

see also http://www.opendoar.org/find.php?format=charts

Content	Explanation
URL	
Domain Name	
Aliases	
Repository Type	• Institutional • Disciplinary • …
Country	
Repository Software	
Dataformat Export Metadata, Download-data	e. g. • SAS • SPSS • EXCEL • CSV (Comma Separeted Values) • …

1.2 Informations about Statistics

Content	Explanation
Statistics Software, Version	e. g. AWStats, Analog
Robot Detection List*	e. g. robots.pm
Configuration*	e. g. awstats.conf
Counting Method**	e. g. Full text in one file: Sum of downloads per month Full text consists of more than one file: • Maximum of file downloads • Sum of file downloads

* Files attached

Sabine Henneberger, Humboldt-Universität zu Berlin, 02/02/2010

**

If a text document consists of more than one part, it must be known, how the monthly number of downloads is counted.

Example: The pdf document consists of more than one part. The number of downloads can be stored

- for every part of the document
- as the sum of the downloads of every part
- as the maximum of downloads of every part

If available, knowing number of Downloads for every part of the document is preferable.

2 Data

The data can be stored in one single table, covering meta data and download data, or in several tables for meta data and download data as well. Meta data can be delivered also in multiple files.

2.1 Necessary Meta Data

Content	**Explanation**
ID	
Date Created	Creation of Meta Data
Classification Content Type (Document Type)	Type of Document: e. g. • Thesis (Bachelor, Master, Dissertation) • Book, Part of Book • Conference Proc., Conference Paper • Report • Journal, Journal Article • Series, Series • Pre-, Postprint • Research Paper • … or like http://www.opendoar.org/find.php?format=charts or like available

Language	
Classification Subject (Subject Area)	e. g. like Subject Area in http://www.opendoar.org/find.php?format=charts and/or DDC (Dewey Decimal Classification) or like available
Classification of Presentation	if documents are represented in special portals: like http://scidok.sulb.uni-saarland.de/sulb/portal/index.php or special collections
Author	Name
Publisher	Name, Institution
Filetype Full Text	XML, HTML, PDF, …
Number Parts Full Text	

2.2 Download Data

Content	**Explanation**
ID	
Type, Filetype	Frontdoor, XML, HTML, PDF, …
Month Access	Month, Year (String, Date) e. g. 'Dec 2008' Time period 2 or more years e.g. 2007 until 2009
Downloads	Number of downloads per month, if available, for every part of full text

Anhang B: Methode in 6 Schritten

Alle Screenshots stammen aus dem Programm SPSS.

Schritt 1: Erstellung des Metadatenfiles

Labels für P_Typ und I_Klasse:

```
VALUE LABELS P_Typ
'A' 'Dissertation'
'B' 'Habilitation'
'C' 'Abschlussarbeit'
'D' 'Festschrift'
'E' 'Einzelpublikation'
'F' 'Komplettpublikation'
'M' 'Missing'.
VARIABLE LABELS P_Typ  'Formaler Publikationstyp'.
MISSING VALUES P_Typ ( 'M' ).


VALUE LABELS I_Klasse
'0' 'Informatik, Informationswissenschaft, allgemeine Werke'
'1' 'Philosophie und Psychologie'
'2' 'Religion'
'3' 'Sozialwissenschaften'
'4' 'Sprache'
'5' 'Naturwissenschaften und Mathematik'
'61' 'Medizin'
'6' 'Technik, angewandte Wissenschaften'
'7' 'Künste und Unterhaltung'
'8' 'Literatur'
'9' 'Geschichte und Geografie'
'M' 'Missing'.
VARIABLE LABELS I_Klasse 'Inhaltsklasse'.
MISSING VALUES I_Klasse ( 'M' ).
```

Schritte 2: Analyse des Metadatenfiles

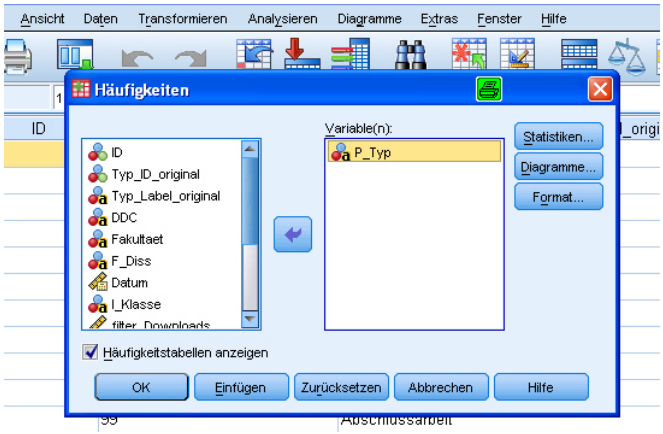


Abb. 94: Screenshot Häufigkeitstabelle

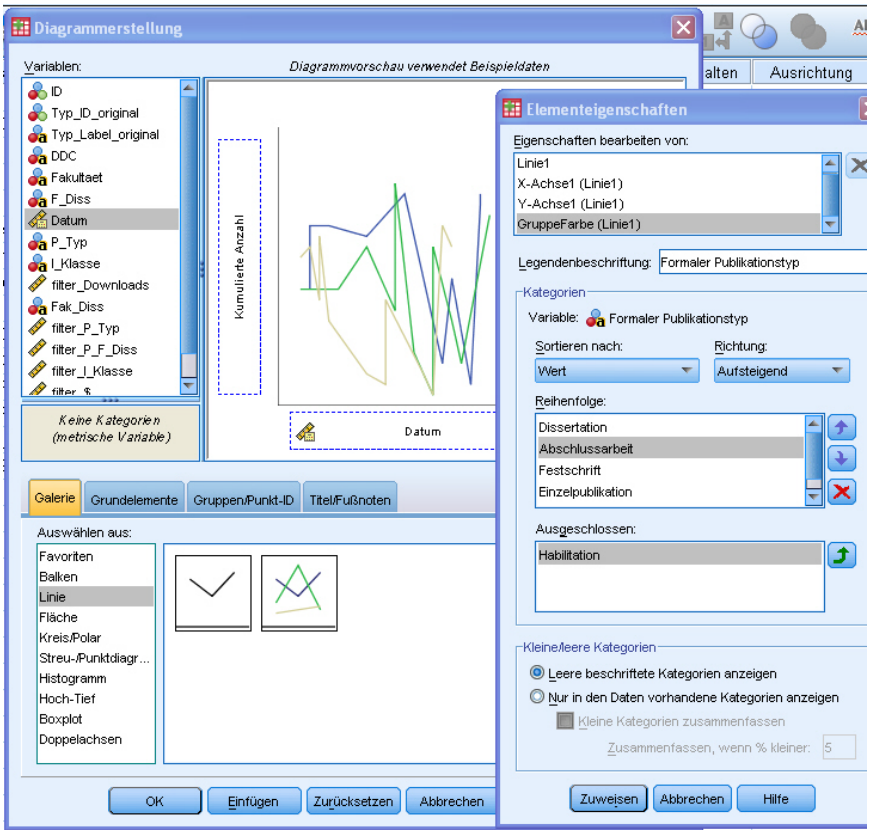


Abb. 95: Screenshot Liniendiagramm, Entwicklung des Bestandes

Schritt 3: Erstellung des Downloadfiles

Auswahl der Publikationen, die mindestens 6 Monate vor dem ersten Erhebungszeitraum online publiziert waren, für HeiDok 2009:

```
COMPUTE filter_Datum = (Datum < DATE.MOYR(8,2009))
```

Schritt 4: Bestimmung der zugelassenen Kategorien

Berechnung des Filters für die Auswahl von formalen Publikationstypen und Fakultäten von Dissertationen zur Analyse:

```
COMPUTE filter_P_Typ=(P_Typ ~= 'B'))
```

Habilitationen (B) werden ausgeschlossen.

```
COMPUTE filter_P_F_Diss=(P_Typ='A' & not(ANY(F_Diss,'7','8','9','12','13')))
```

Ausgewählt werden Dissertationen (A) und die rot markierten Fakultäten werden ausgeschlossen.

Schritt 5: Signifikanztest für zugelassene Kategorien

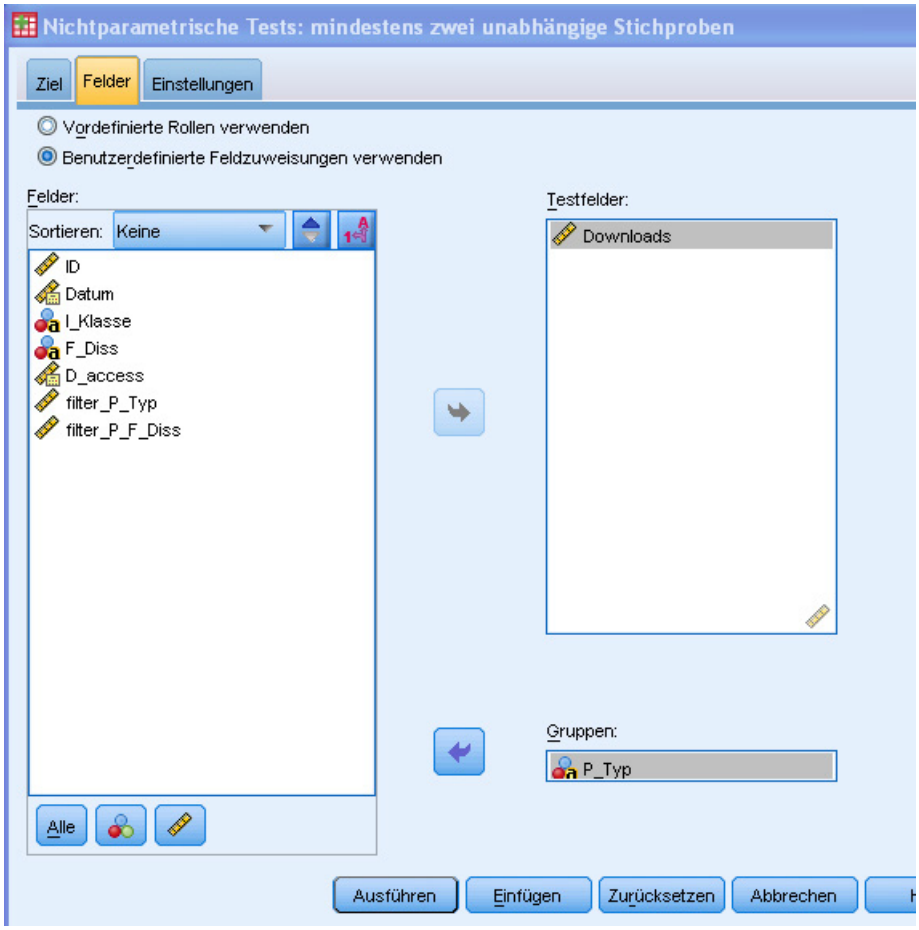


Abb. 96: Screenshot Signifikanztest, Auswahl der Test- und Gruppenvariablen

Anweisung zur Zusammenfassung von Kategorien des Merkmals F_Diss:

```
STRING F_Diss_Sign (A2).
DO IF Any(F_Diss, '3','1','2').
COMPUTE F_Diss_Sign = 'M1'.
ELSE IF Any(F_Diss, '5','10','11').
COMPUTE F_Diss_Sign = 'M2'.
ELSE.
COMPUTE F_Diss_Sign = F_Diss.
END IF.
EXECUTE.
```

Schritt 6: Grafische Darstellung der Ergebnisse

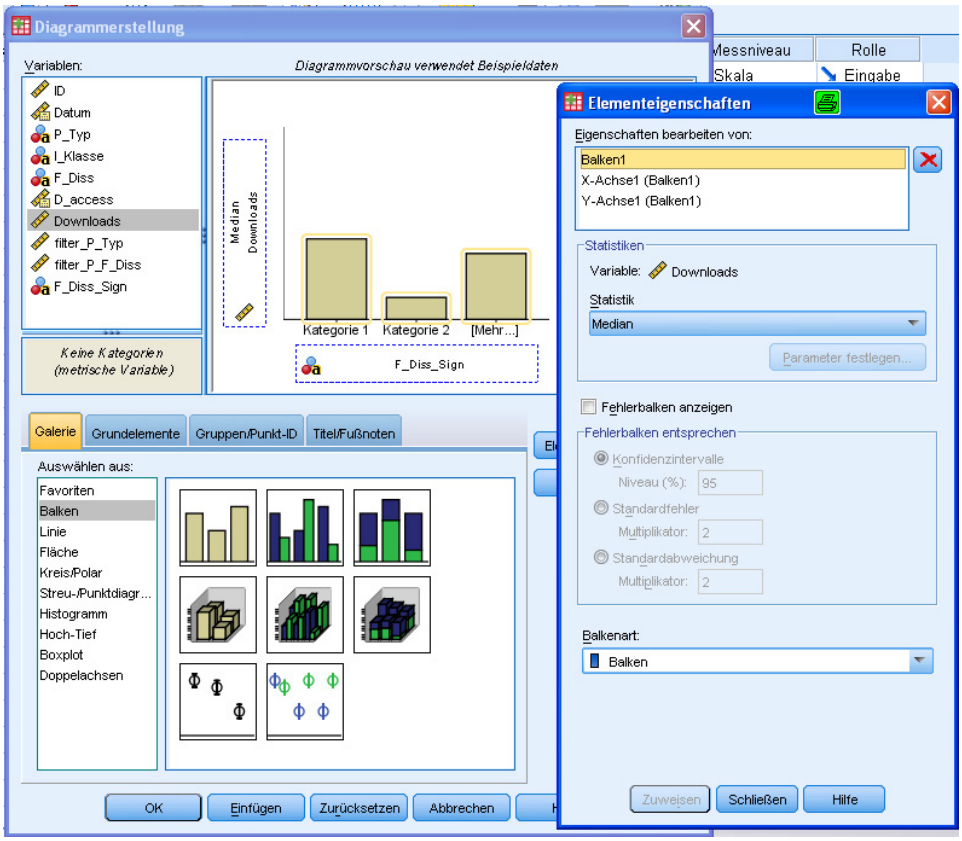


Abb. 97: Screenshot Balkendiagramm, Vergleich der Mediane